# Handleiding voor de ontwikkeling van taaltoetsen

## Te gebruiken met het ERK

**Samengesteld door ALTE, in opdracht van de Language Policy Division, Raad van Europa**

ALTE Association of Language Testers in Europe

taal: unie | CNaVT Certificaat Nederlands als Vreemde Taal

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

# Handleiding voor de ontwikkeling van taaltoetsen

Te gebruiken met het ERK

Samengesteld door ALTE in opdracht van de Language Policy Division, Raad van Europa

Vertaald door het CNaVT
Centrum voor Taal en Onderwijs (KU Leuven)

Originele titel: *Manual for Language Test Development and Examining* – © Raad van Europa, 2011

© Nederlandse uitgave: Certificaat Nederlands als Vreemde Taal (CNaVT), 2017
Coverfoto: newco500 / 123RF Stockfoto
ISBN: 9789079219094

# Inhoud

# Voorwoord

Deze handleiding is een welkome en broodnodige aanvulling op de 'toolkit' die ondersteuning biedt aan specifieke groepen die gebruik willen maken van het Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen (ERK). Wij willen de *Association of Language Testers in Europe* (ALTE) bedanken omdat zij, nadat de Raad van Europa had gevraagd dit document voor te bereiden en in overeenstemming met haar status van ondersteunende Niet-Gouvernementele Organisatie (NGO) binnen de Raad van Europa, opnieuw een fundamentele bijdrage heeft geleverd aan het efficiënte gebruik van het ERK.

Het ERK werd ontwikkeld met de bedoeling een gemeenschappelijke basis te creëren waarover verschillende partners uit het veld konden discussiëren en communiceren, over de grenzen heen van de Europese landen die lid zijn van de Raad van Europa, binnen de lerarenopleidingen, bij het ontwikkelen van taalcursussen, leerplannen, handboeken, toetsen, etc. Het wordt aan de gebruikers aangeboden als een beschrijvend instrument dat hen helpt na te denken over hun beslissingen en handelingen, en hun inspanningen kadert en coördineert. Op die manier helpt het taalleerders in hun specifieke contexten vooruit. Het ERK is dus een flexibel instrument dat aan specifieke gebruikscontexten moet worden aangepast. Dit is fundamenteel. Dit idee wordt dan ook weerspiegeld in het niveausysteem zelf, dat flexibel kan worden gebruikt en aangepast naargelang het soort leer- , onderwijs- en evaluatiedoelen dat wordt ontwikkeld, zoals descriptoren voor verschillende niveaus in bepaalde talen en contexten.

De descriptoren, die 'transparant, bruikbaar en relevant bevonden werden door groepen taaldocenten/ leerkrachten (moedertaalsprekers en niet-moedertaalsprekers) uit uiteenlopende onderwijssectoren met heel verschillende opleidings- en ervaringsprofielen' (ERK, p.30), zijn niet altijd allesomvattend en zeker niet normatief bedoeld. Gebruikers worden uitgenodigd om ze te gebruiken, aan te passen of te vervolledigen naargelang de context en behoefte. Deze praktische handleiding is een waardevolle leidraad voor wie op deze manier taalvaardigheidstoetsen wil ontwikkelen, gelinkt aan de ERK-niveaus, op een principiële en niet dwingende manier.

Van iedereen die taalcursussen aanbiedt, wordt kwaliteit, samenhang en duidelijkheid verwacht. Daarenboven wordt het alsmaar belangrijker dat bewijzen van taalvaardigheid in verschillende geografische gebieden erkend worden. Daarom krijgt het ERK alsmaar meer aandacht en worden de ERK-niveaus niet alleen in Europa, maar ook ver daarbuiten gebruikt als referentiepunt en ijkinstrument. Wij zijn hier heel blij mee, maar we willen gebruikers zeker ook aanmoedigen om op ontdekking te gaan en ervaringen te delen over hoe het ERK in al zijn dimensies verder kan gebruikt worden om taalleerders (die allen heel verschillend zijn) te ondersteunen en hun dynamische proces van levenslang leren te erkennen. Het is uiteindelijk aan hen om hun leerproces te plannen en te evalueren in het licht van hun behoeften en de veranderende contexten waarin ze zich bevinden. De initiatieven van de Raad van Europa om meertalig en intercultureel onderwijs en een globale aanpak van alle talen in het onderwijs te promoten, houdt nieuwe uitdagingen in - uitdagingen voor de ontwikkeling van leerplannen, voor de manier waarop er lesgegeven en geëvalueerd wordt. We kijken uit naar de essentiële bijdragen die organisaties van deskundigen zoals ALTE leveren om de waarden van de Raad van Europa te helpen promoten in het domein van het taalonderwijs.

**Joseph Sheils**
*Language Policy Division*
*Raad van Europa*

# Inleiding

## Achtergrond

In 2001 werd de definitieve versie van het Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen (ERK) gepubliceerd, en sindsdien is de interesse voor dit document alleen maar toegenomen, niet alleen in Europa, maar wereldwijd. De impact van het ERK is veel groter dan verwacht en er bestaat geen twijfel over dat het referentiekader gezorgd heeft voor een grotere bewustwording rond belangrijke facetten van het leren, onderwijzen en evalueren van talen. De Raad van Europa heeft ook bijgedragen tot de ontwikkeling van een 'toolkit' die bestaat uit bronnen die meer informatie verschaffen over het ontstaan en het gebruik van het ERK voor zowel beleidsmakers, leerkrachten, toetstontwikkelaars en andere betrokkenen.

Zoals Daniel Coste (2007), een van de auteurs van het ERK, opmerkt, is de invloed van het ERK op taaltoetsing opvallend. Er is binnen het domein van taaltoetsen vooral aandacht gegaan naar een welbepaald aspect van het referentiekader, namelijk het relateren van toetsen aan het ERK. Momenteel zijn er een aantal nuttige hulpmiddelen uit de toolkit beschikbaar. Deze omvatten:

- *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Raad van Europa, 2009);
- een technisch *Reference Supplement to the Manual for Relating Examinations to the CEFR* (Banerjee, 2004; Verhelst, 2004 a,b,c,d; Kaftandjieva, 2004; Eckes, 2009);
- voorbeeldmaterialen die de ERK-niveaus illustreren;
- een hulpmiddel voor de inhoudsanalyse van materialen om te spreken, schrijven, luisteren en lezen;
- DESCRIPTOREN voor de verschillende niveaus in het Engels en andere talen.

De Raad van Europa heeft ook gelegenheden gecreëerd (*Reflections on the Use of the Draft Manual for Relating Language Examinations to the CEFR*, Cambridge, 2007; werkgroep voor de conferentie, EALTA-conferentie, Athene, 2008) voor gebruikers om van gedachten te wisselen over het gebruik van de hierboven vermeldde handleiding en hun ervaringen met de verschillende afstemmingssfases die erin besproken worden.

*The Association of Language Testers in Europe* (ALTE) heeft – als internationale niet-gouvernementele organisatie met adviserende rol binnen de Raad van Europa – bijgedragen aan de bronnen die deel uitmaken van de toolkit, zoals het EAQUALS/ALTE European Portfolio (ELP) en de *ALTE Content Analysis Grids*. ALTE werd ook vertegenwoordigd door Dr. Piet van Avermaet in het auteursteam die de *Manual for Relating Language Examinations to the CEFR* vormgaf. Samen met de *Language Policy Division* van de Raad van Europa, wil ALTE de gebruikers van de toolkit graag aanmoedigen om het ERK op een efficiënte manier te hanteren in hun eigen context, en om hun eigen doelstellingen te bereiken.

## Waarom deze handleiding

De handleiding *The Manual for Relating Language Examinations to the CEFR* die hierboven ter sprake kwam, werd speciaal ontwikkeld om te helpen bij het relateren van toetsen aan het ERK. Samen met het *Reference Supplement* behandelt het document een algemene aanpak en schetst het opties, zoals een cesuurbepaling.

Deze *Handleiding voor de ontwikkeling van taaltoetsen* is een aanvulling. Het legt de nadruk op die aspecten van toetsontwikkeling die niet behandeld werden in de andere handleiding. Eigenlijk is het een bewerking van een document dat de Raad van Europa eerder uitgaf en bekend stond onder de naam *Users' Guide for Examiners* (1996), een document dat de Raad van Europa samen met andere gebruikershandleidingen liet schrijven om de eerste kladversie van het ERK te begeleiden in 1996/7.

ALTE was verantwoordelijk voor de productie van de originele versie. Het laatste decennium waren er heel wat ontwikkelingen in de validiteitstheorie en nam het gebruik en de impact van het ERK toe. Daarom was het tijd om het document grondig te herwerken. ALTE was blij met de vraag om de herwerkingen te coördineren, en vele individuele ALTE-leden droegen dan ook bij aan dit proces.

Tijdens de herwerkingen was het nuttig om onszelf te herinneren aan de grondslag en de doelen van het ERK, en om die doelen een duidelijke plaats te geven in deze handleiding voor potentiële gebruikers.

Het ERK was als gemeenschappelijk referentiekader eerst en vooral bedoeld als een instrument om aan te zetten tot '*denken, communiceren en meer zelfstandigheid*' (Trim, 2010). Het werd ontwikkeld om een gemeenschappelijke taal te bieden die zou helpen om de communicatie te vergemakkelijken in de domeinen van talen leren, onderwijzen en evalueren. Het ERK voorziet ook in een aantal referentieniveaus om de maat van taalvaardigheid te meten, gaande van een beginnersniveau (A1) tot een heel gevorderd niveau (C2), rekening houdende met verschillende vaardigheden en gebruiksdomeinen.

Daarom is het een geschikt instrument om praktijkvoorbeelden uit heel verschillende contexten uit Europa en ver daarbuiten met elkaar te vergelijken. Hier moet wel aan toegevoegd worden dat dit gemeenschappelijk referentie-instrument toch niet altijd en overal gebruikt kan worden zonder enige aanpassing door de gebruiker aan zijn lokale context en doelstellingen.

De auteurs van het ERK hebben hier de nodige nadruk op gelegd. In de inleidende tekst voor de gebruiker bijvoorbeeld, schrijven ze: *'We have NOT set out to tell practicioners what to do or how to do it'* (p. ix), en ze herhalen dit standpunt ook doorheen de rest van de tekst. Latere bronnen die deel zouden uitmaken van de toolkit zoals de *Manual for Relating Language Examinations to the CEFR*, lagen in dezelfde lijn. De auteurs van de handleiding wijzen erop dat het niet de enige manier is om toetsen op het ERK af te stemmen en dat geen enkele instelling verplicht is om de link ook maar te leggen (p. 1).

Op het beleidsforum van de Raad van Europa dat in 2007 in Straatsburg werd gehouden, werd het gebruik van het ERK besproken en toen merkte Coste op hoe contextueel gebruik, dat gezien wordt als doelbewuste interventie, *'can take various forms, apply on different levels, have different aims, and involve different types of players.'* Hij voegt daaraan toe: '*All of these many contextual applications are legitimate and meaningful but, just as the Framework itself offer a range of (as it were) built-in options, so some of the contextual applications exploit it more fully, while others extend or transcend it.*' Met andere woorden, als er beslissingen moeten vallen over het afstemmen van een toets dan is het belangrijk te onthouden dat het ERK niet bedoeld is als maatstaf en dat er meer dan één manier bestaat om een toets af te stemmen op de specifieke context en doeleinden.

Zoals Jones en Saville (2009, p. 54-55) schrijven:

> *'…some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way, we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution.'*

De *Manual for Relating Language Examinations to the CEFR* focust op '*procedures involved in the justification of a claim that a certain examination or test is linked to the CEFR*' en '*does not provide a general guide how to construct good language tests or examinations*' (p. 2). Het is dus geen handleiding om een goede toets te maken, maar eerder een gids die helpt bij het leggen van een link tussen een toets en het ERK. De aanvullende aanpak die in deze handleiding wordt besproken, start vanaf het ontwikkelen van de toets zelf en toont hoe er een link met het ERK kan gemaakt worden tijdens elke stap in het proces, zodat:

- de inhoud van de toets kan worden bepaald;
- specifieke taalvaardigheidsniveaus kunnen worden bepaald;
- de productie van gesproken en geschreven taal op een taaltoets kan worden geïnterpreteerd, op een manier die steek houdt in de echte wereld.

Daarom is deze handleiding ruimer dan de drie belangrijke manieren van gebruik[1] die in het ERK (Nederlandse Taalunie, 2008) zelf worden besproken, namelijk:

1. voor de specificatie van de inhoud van toetsen: *wat wordt getoetst*
2. voor het benoemen van de criteria op grond waarvan men bepaalt of een doelstelling is verwezenlijkt: *hoe worden prestaties beoordeeld*

1 Noot van de vertaalster: de drie belangrijke manieren waarop het Referentiekader kan worden gebruikt, werden niet vertaald uit het origineel, maar letterlijk overgenomen uit het ERK, p. 159 en 160.

3. het beschrijven van niveaus van taalvaardigheid in bestaande toetsen, waardoor vergelijkingen mogelijk worden tussen verschillende kwalificatiesystemen: *hoe vergelijkingen kunnen worden gemaakt*

Deze handleiding biedt een samenhangende gids voor toetsontwikkeling in het algemeen en is nuttig voor uiteenlopende doeleinden. Toetsontwikkeling wordt erin behandeld als een cyclus want succes in een bepaalde fase hangt samen met het werk dat in een vorige fase werd verricht. De hele cyclus moet efficiënt verlopen zodat elke fase goed werkt. Hoofstukken 1 t.e.m. 5 geven een overzicht van de cyclus en behandelen elke fase meer in detail:

**Hoofdstuk 1** – fundamentele begrippen – introduceert de basisconcepten van taalvaardigheid, namelijk VALIDITEIT, BETROUWBAARHEID en rechtvaardigheid

**Hoofdstuk 2** – een toets ontwikkelen – gaat van de beslissing om een toets aan te bieden tot het ontwikkelen van definitieve TOETSSPECIFICATIES

**Hoofdstuk 3** – een toets samenstellen – beschrijft hoe ITEMS worden geschreven en toetsen worden samengesteld

**Hoofdstuk 4** – een toets afleveren – beschrijft de AFNAME van toetsen, vanaf de registratie van de kandidaten tot het terugsturen van de de toetsmaterialen naar de toetsinstelling

**Hoofdstuk 5** – corrigeren, scores toekennen en resultaten rapporteren– vervolledigt de operationele cyclus

**Hoofdstuk 6** – monitoring en controle – beschrijft hoe de cyclus kan worden herhaald over de jaren heen zodat de kwaliteit en het nut van de toets kunnen worden verbeterd.

## De lezer van deze handleiding

Deze handleiding is bedoeld voor iedereen die geïnteresseerd is in het ontwikkelen en gebruiken van taaltoetsen die aan het ERK gelinkt zijn. Het document werd geschreven voor TOETSONTWIKKELAARS die nieuw zijn in het veld, maar evengoed voor meer ervaren gebruikers. Het introduceert namelijk de gemeenschappelijke principes die normaal worden toegepast, zowel door grote toetsinstellingen die toetsen ontwikkelen voor duizenden kandidaten op verschillende locaties, als door individuele leerkrachten die hun studenten willen toetsen in hun eigen klaslokaal. De principes zijn dezelfde voor toetsen die een grote impact hebben op de toekomst van de kandidaat, als voor toetsen die helemaal vrijblijvend zijn. De praktische stappen die worden ondernomen kunnen verschillend zijn, maar veranderen hier niks aan.

We gaan ervan uit dat de lezers van deze handleiding het ERK kennen, of bereid zijn het te gebruiken terwijl ze deze handleiding lezen.

## Gebruik van de handleiding

De principes van taaltoetsing die hier worden besproken, zijn vrij algemeen toe te passen, maar de aanbieder van een toets moet beslissen hoe hij deze toepast in zijn eigen context. In de handleiding staan een aantal voorbeelden en tips over het uitvoeren van bepaalde activiteiten. Vanzelfsprekend zal dit praktische advies op sommige contexten meer van toepassing zijn dan op andere, afhankelijk van het doel van de toets en de middelen waarover de ontwikkelaars beschikken. Toch is het daarom niet minder geschikt voor bepaalde lezers: als de gebruikers de principes begrijpen, dan kunnen zij dankzij de voorbeelden nadenken over hoe ze die principes kunnen toepassen in hun eigen context.

Los van het ERK bestaan er nog andere bronnen die kunnen helpen om toetsen af te stemmen op het ERK en deze handleiding is slechts een van de bronnen die samen de toolkit vormen die door de Raad van Europa werden gemaakt en ter beschikking gesteld. Daarom wordt er ook geen informatie of theoretische kennis gegeven die makkelijk ergens anders te vinden is. Zoals gezegd, probeert deze handleiding geen informatie te herhalen die al in de *Manual for Relating Language Examination to the CEFR* staat, maar wil het net een aanvulling zijn op deze publicatie.

Deze handleiding hoeft niet van a tot z worden gelezen. Als verschillende taken tijdens de ontwikkeling en praktische organisatie van toetsen door verschillende mensen worden uitgevoerd, dan kan elke persoon

dat stuk lezen dat voor hem relevant is. Niettemin is het nuttig, zelfs voor wie in een bepaald onderdeel gespecialiseerd is, om de volledige toetscyclus in deze handleiding van naderbij te bekijken.

Voor wie meer wil lezen, staat er achter elk hoofdstuk aanbevolen literatuur die de onderwerpen uit deze handleiding uitgebreider behandelt of meer praktische handvaten aanreikt. Deze bronnen worden telkens voorafgegaan door een aantal kernvragen die de lezer kan gebruiken om na te gaan of hij het gelezen hoofdstuk volledig begrepen heeft.

Deze handleiding is niet normerend, maar belicht enkel de belangrijkste principes en benaderingen van toetsontwikkeling en evaluatie waarnaar de gebruiker kan verwijzen wanneer hij toetsen ontwikkelt voor zijn eigen gebruikscontext. In de handleiding staan geen recepten om toetsvragen te maken gebaseerd op de verschillende schalen van het ERK, want hoewel ze gedetailleerd genoeg zijn om een gemeenschappelijk referentiekader te vormen, werden ze niet ontwikkeld om precieze vergelijkingen te maken.

Inderdaad, in een van de eerste kladversies van het referentiekader (Straatsburg, 1998) waren de voorbeeldschalen opgenomen als bijlage en werden ze niet in de tekst zelf vermeld. De *Common Reference Levels* waren de enige schalen in de tekst. De originele opmaak van de kladversie uit 1998 onderstreepte het verschil in status en functies van de algemene referentieniveaus en de meer specifieke voorbeeldschalen. Zo bleek nog maar eens dat het om schalen onder voorbehoud ging; sommige waren niet eens geijkt of kwamen op de C-niveaus heel weinig voor.

In de kladversie van het ERK uit 1998 werd deze indicatieve status van de voorbeeldschalen expliciet in de tekst vermeld (p. 131):

> *'The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organize or describe their system of levels and modules. It is also to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time as the experience of member states and of institutions with related expertise is incorporated into the description.'*

Als de schalen op een overdreven normerende wijze gebruikt zouden worden, bestaat het gevaar dat er een generieke benadering ontstaat om de taalvaardigheid te meten. De functionele en linguïstische schalen zijn bedoeld om de brede toepasbaarheid van de niveaus te illustreren en niet zozeer om deze gedetailleerd en nauwkeurig te beschrijven. Gezien de vele verschillende demografische contexten, doeleinden, onderwijs- en leerstijlen is het immers onmogelijk om een 'typische B1' student te omschrijven bijvoorbeeld. Een logisch gevolg is dat het moeilijk is om een syllabus of toets te ontwikkelen voor B1, of voor eender welk ander niveau, die gepast is voor alle contexten.

Wil het ERK een duurzaam en positief effect hebben, dan moeten de principes en voorbeelden uit dit document geïntegreerd worden in de routineprocedures van toetsinstellingen. Zo kunnen er over de jaren heen deskundige systemen worden ontwikkeld om de afstemming op het ERK steeds beter te onderbouwen. Dit houdt ook in dat de tekst van het ERK in zijn totaliteit wordt gebruikt en waar nodig aangepast wordt aan specifieke contexten en toepassingen.

Omdat het niet mogelijk is om aan de hand van één enkele cesuurbepaling een stabiel en consistent bewijs te leveren voor de afstemming van een toets op het ERK, is het belangrijk dat toetsinstellingen doorheen de tijd bewijzen opbouwen. Dit betekent dat de aanbevelingen uit de *Manual for Relating Language Examinations to the CEFR* en andere bronnen uit de toolkit die gebruikt worden voor de afstemming op het ERK geïntegreerd zouden moeten worden in standaardiseringsprocedures van de toetsinstellingen en niet als een eenmalig iets zouden mogen worden gezien.

Daartoe wil deze handleiding de lezer aansporen. Het onderstreept hoe belangrijk het is om systemen te ontwikkelen en te onderhouden die toelaten om standaarden te bepalen en te controleren, door de jaren heen.

## Conventies in de handleiding

In deze handleiding worden de volgende conventies gehanteerd:

- Er wordt naar de *Manual for Relating Language Examination to the CEFR* verwezen als 'handleiding'.
- Naar het *Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen* wordt verwezen met 'ERK'.
- De organisatie die verantwoordelijk is voor het maken van een toets wordt de 'toetsinstelling' genoemd. Termen zoals 'toetsontwikkelaar' worden sporadisch gebruikt om te verwijzen naar personen die een specifieke functie hebben binnen de ontwikkelingscyclus.
- Woorden die in de Verklarende woordenlijst staan (Bijlage VIII), worden benadrukt met KLEINKAPITALEN wanneer ze de eerste keer worden vermeld in deze handleiding en op andere plaatsen waar dit de lezer helpt.

**Dr Michael Milanovic**
*ALTE-manager*

# 1 Fundamentele begrippen

De praktische richtlijnen in deze handleiding om taaltoetsen te maken hebben een gefundeerde basis nodig van onderliggende principes en theorieën. Dit hoofdstuk licht toe:

- wat taalvaardigheid is;
- waarom validiteit een essentiële eigenschap van een goede toets is;
- wat betrouwbaarheid is;
- het concept van rechtvaardigheid in toetsen.

Dit deel beschrijft ook het ontwikkelingsproces van een toets. Dit proces wordt in de andere hoofdstukken verder uitgelegd.

## 1.1 Hoe wordt taalvaardigheid bepaald?

### 1.1.1 Modellen voor taalgebruik en –vaardigheden

Taal gebruiken is een heel complex proces dat beroep doet op een groot aantal vaardigheden. Het is belangrijk om bij de start van een toetsproject een expliciet model voor deze vaardigheden uit te werken waarin ook opgenomen wordt hoe de verschillende vaardigheden zich tot elkaar verhouden. Dit model hoeft niet te argumenteren hoe taalvaardigheid in onze hersenen werkt, maar het kan dit wel doen. Het is bedoeld om ons te wijzen op opmerkelijke aspecten van vaardigheden zodat we die verder kunnen in overweging nemen. Het is het uitgangspunt om te beslissen welke aspecten van vaardigheden er kunnen of zouden moeten getoetst worden. Het helpt ervoor te zorgen dat de toetsresultaten kunnen worden geïnterpreteerd en zinvol zijn. De mentale eigenschap die door het model zichtbaar wordt, wordt ook het CONSTRUCT genoemd. Een construct is dus een theoretsiche voorstelling van een abstract proces.

### 1.1.2 Het ERK-model van taalgebruik

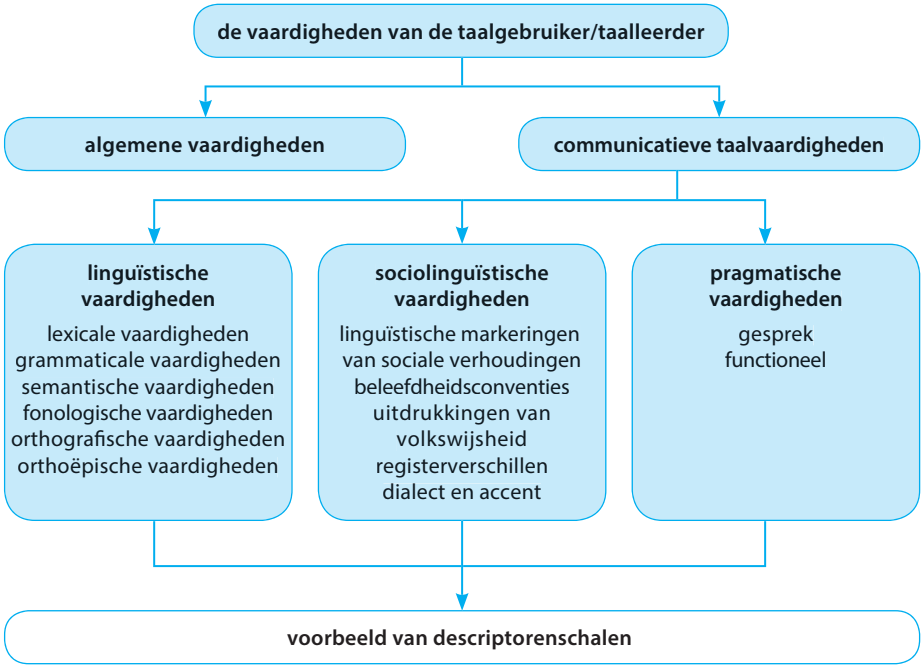
Verschillende auteurs hebben invloedrijke modellen van taalvaardigheid voorgesteld (bv. Bachman, 1990; Canale en Swain, 1981; Weir, 2005).

Het is zinvol om deze handleiding te starten met het model uit het ERK voor taalgebruik en –leren. Deze 'ACTIEGERICHTE AANPAK' wordt beschreven als:

> *'…the actions performed by persons who as individuals and as social agents develop a range of* ***competences****, both* ***general*** *and in particular* ***communicative language competences****. They draw on the competences at their disposal in various contexts under various* ***conditions*** *and* ***constraints*** *to engage in* ***language activities*** *involving* ***language processes*** *to produce and/or receive* ***texts*** *in relation to* ***themes*** *in specific* ***domains****, activating those* ***strategies*** *which seem most appropriate for carrying out the* ***tasks*** *to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences' (CEFR p. 9 (emphasis in original)).*

In deze paragraaf worden de belangrijkste bouwstenen uit het model zichtbaar die in de tekst van het ERK in detail worden besproken. De hiërarchische organisatie van de bouwstenen wordt weergegeven in de titels en ondertitels van hoofdstuk 4 en 5 van het ERK.

Figuur 1 illustreert dit aan de hand van enkele titels en ondertitels uit hoofdstuk 5, De vaardigheden van de taalgebruiker/taalleerder. In de figuur zijn de vaardigheden onderverdeeld in twee subgroepen: de Algemene vaardigheden (zoals de Declaratieve kennis en Existentiële vaardigheden, hier niet opgenomen) en de Communicatieve taalvaardigheden, die verder onderverdeeld worden in drie groepen: Linguïstische, Sociolinguïstische en Pragmatische vaardigheden. Elke vaardigheid wordt daarna nog verder opgesplitst.




**Figuur 1** ***Een beperkt overzicht van Hoofdstuk 5 van het ERK: De vaardigheden van de taalgebruiker/taalleerder***

Hieraan gerelateerd is hoofdstuk 4, waarin de communicatieve doelen en manieren van taalgebruik worden besproken. Zoals figuur 2 aangeeft, betekent dit dat het duidelijk moet zijn *wat* er gecommuniceerd wordt (thema's, taken en doelen), maar dat er ook moet nagedacht worden over de activiteiten die plaatsvinden en de strategieën die gehanteerd worden, en dus over de functionele taalvaardigheden die leerders gebruiken wanneer ze communiceren. Voor de duidelijkheid wordt er in figuur 2 slechts een deel van deze complexe hiërarchie getoond.


**Figuur 2** ***Een beperkt overzicht van Hoofdstuk 4 van het ERK: Taalgebruik en de taalgebruiker/taalleerder***

### 1.1.3 Operationalisering van het model

Als we ons afvragen hoe we het MODEL VAN TAALGEBRUIK kunnen operationaliseren, dan moeten we twee belangrijke aspecten in het achterhoofd houden die voor een groot stuk zullen bepalen hoe onze toets er zal uitzien: de AUTHENTICITEIT van de ITEMS en TAKEN en de mate waarin de vaardigheden afzonderlijk van elkaar getoetst worden.

**Authenticiteit**

Twee belangrijke aspecten van authenticiteit in taaltoetsen zijn de *situationele* en *interactieve* authenticiteit. Met *situationele authenticiteit* wordt verwezen naar de representativiteit van taken en items voor taalactiviteiten uit het echte leven. Interactieve authenticiteit daarentegen heeft betrekking op de natuurlijkheid van de interactie tussen de kandidaat en de taak en op de mentale processen die ermee gepaard gaan. Een taakgerichte luistertaak waarbij er specifieke informatie moet worden gevonden, kan situationeel authentieker worden als er een alledaagse context wordt gecreëerd zoals een weersvoorspelling. Dezelfde taak kan interactief authentieker worden als de kandidaat een reden krijgt waarom hij moet luisteren, bv. omdat hij een picknick moet plannen die week en een geschikte dag moet uitkiezen.

In taaltoetsen moeten we vaak verschillende soorten authenticiteit in balans brengen om een geschikte taak te creëren. We moeten bv. materialen en activiteiten aanpassen aan het taalniveau dat de kandidaat op een bepaald moment heeft in de doeltaal. Dit maatwerk betekent dat de interactie die de kandidaten met de teksten en met elkaar aangaan authentiek is, hoewel de materialen dat misschien niet helemaal zijn.

Om een taak situationeel authentieker te maken, moet er eerst bepaald worden wat de taak in het echte leven maakt tot wat ze is. Die elementen moeten zo nauwkeurig mogelijk worden gekopieerd. De interactieve authenticiteit kan op verschillende manieren verhoogd worden:

- Situaties en taken gebruiken die met grote zekerheid gekend en relevant zijn voor zowel de kandidaat als het niveau.
- Duidelijk maken met welk doel een bepaalde taak moet worden uitgevoerd, wie de ontvangers van de boodschap zijn, en in welke context alles zich afspeelt.
- Duidelijk maken hoe de taak succesvol kan worden uitgevoerd.

**Vaardigheden integreren**

Voor wie een model voor taalgebruik, of construct, ontwerpt, kan het theoretisch interessant zijn om een onderscheid te maken tussen verschillende vaardigheden. Toch is het ontzettend moeilijk om in authentieke taken verschillende vaardigheden van elkaar te scheiden. Dit komt doordat we in communicatieve situaties vaak meerdere vaardigheden inzetten op hetzelfde moment. Als een taalleerder iemand probeert te begrijpen die hem op straat de weg vraagt, worden er een aantal vaardigheden aangesproken: grammaticale en tekstuele vaardigheden om de boodschap te ontcijferen, sociolinguïstische vaardigheden om de sociale communicatiecontext te begrijpen en illocutionaire vaardigheden om te interpreteren wat de spreker wil bereiken.

Als we een toetstaak willen ontwikkelen, dan is het belangrijk om een evenwicht te vinden tussen de vaardigheden die nodig zijn voor een succesvolle *RESPONS* (antwoord). Sommige zullen belangrijker zijn dan andere – zij zullen de focus vormen van de taak. De taak zou genoeg gepaste taal moeten ontlokken zodat er een oordeel kan geveld worden over de bekwaamheid van de kandidaat in de gekozen vaardigheid (vaardigheden). Het is ook belangrijk om na te denken over de manier waarop de respons GECORRIGEERD wordt en hoe er punten worden aan toegekend (zie Hoofdstuk 2.5 en 5.1.3): de beheersing van een bepaalde vaardigheid zou de basis moeten vormen voor de beoordeling.

### 1.1.4 De niveaus van het ERK

Als aanvulling op het model dat hierboven werd voorgesteld, onderscheidt het ERK ook zes niveaus van communicatieve taalvaardigheid die als hulp kunnen dienen bij het bepalen van leerdoelen, de evaluatie van leerprocessen of taalvaardigheidsniveaus. Dit conceptueel referentiekader wordt geïllustreerd met descriptorenschalen, geformuleerd in 'can do'-descriptoren.

Een voorbeeld van een 'can do'-descriptor voor een laag leesniveau (A1) is:

*Kan zeer korte, eenvoudige teksten frase voor frase begrijpen door vertrouwde namen, woorden, en elementaire combinates te herkennen en indien nodig te herlezen.*

Vergelijk dit met een descriptor voor een hoog niveau (C2):

*Kan vrijwel all vormen van geschreven taal begrijpen en kritisch interpreteren, met inbegrip van abstracte, structureel complexe of zeer spreektalige literaire en niet-literaire geschriften. Kan een breed scala van lange, complexe teksten begrijpen en daarbij subtiele verschillen in stijl en impliciete en expliciete betekenissen opmerken.*

Raad van Europa (2008, p. 66)

De zes vaardigheidsniveaus worden als volgt benoemd:

| | | |
|---|---|---|
| C2 | Mastery ('beheersing') | } Vaardige gebruiker |
| C1 | Effective Operational Proficiency ('effectieve operationele vaardigheid') | |
| B2 | Vantage ('uitzicht') | } Onafhankelijke gebruiker |
| B1 | Threshold ('drempel') | |
| A2 | Waystage ('tussenstap') | } Basisgebruiker |
| A1 | Breakthrough ('doorbraak') | |

Als taaltoetsers zouden we de 'can do'-descriptoren correct moeten begrijpen. Deze descriptoren zijn illustratief, ze zijn dus:

- niet exhaustief;
- niet prescriptief;
- geen definitie;
- geen leerprogramma;
- geen checklist.

De 'can do'-descriptoren zijn een richtlijn voor wie onderwijs geeft zodat vaardigheidsniveaus kunnen worden herkend, en zodat erover kan worden gepraat. We kunnen ze als een richtlijn gebruiken om toetsen te ontwikkelen, maar we mogen er niet van uitgaan dat we daardoor geen tijd meer moeten investeren in het definiëren van vaardigheidsniveaus.

Toetsontwikkelaars moeten kiezen welke 'can do'-descriptoren het relevantst zijn in hun context, in het DOMEIN van hun toets. Bijvoorbeeld, om het personeel van een hotel te onderwijzen en te toetsen, zouden de descriptoren voor een 'Doelgerichte samenwerking' (ERK, 4.4.3.1) bruikbaar kunnen zijn, maar die voor 'TV en films kijken' waarschijnlijk niet (ERK, 4.4.2.3). Als de bestaande illustratieve SCHALEN of andere materialen in de ERK-toolkit niet nauw genoeg aangepast zijn aan de context, dan kunnen ze aangevuld worden met 'can do'-descriptoren uit andere bronnen, of er kunnen er specifiek voor de context worden geschreven.

**Een toets afstemmen op het ERK**

Het afstemmen van een toets op het ERK begint met het ERK aan te passen aan de context van de toets. Het ERK is immers *'context-free in order to accomodate generalisable results from different specific contexts'*, en is tegelijkertijd '*context-relevant, relatable to or translatable into each and every relevant context*'. (CEFR, p. 21).

Een toets afstemmen op het ERK mag zeker niet betekenen dat er geprobeerd wordt om het ERK in om het even welke context op dezelfde rigide manier toe te passen. De toetsontwikkelaar moet kunnen verantwoorden hoe hij het ERK heeft 'vertaald' naar zijn context, o.a. door die context te specifiëren.

Andere belangrijke contextuele kenmerken zijn die van de kandidaten. Er zijn bijvoorbeeld opmerkelijke verschillen tussen leerders wat betreft hun leeftijd en cognitieve ontwikkeling, de reden waarom ze een taal leren, enz. Aan de hand van enkele van deze verschillen kunnen groepen met heel specifieke eigenschappen worden onderscheiden. Taaltoetsontwikkelaars hebben vaak een welbepaalde groep voor ogen, zoals jonge leerders of volwassenen. Beide groepen kunnen gerelateerd worden aan het ERK, maar het B1-niveau van een jonge leerder verschilt toch van het B1-niveau van een volwassene omdat er verschillende descriptoren mee gemoeid zijn.

De vaardigheidsprofielen van verschillende leerders verschillen vaak van elkaar (sommigen zijn betere luisteraars dan lezers en vice versa). Dit maakt het moeilijk om ze op eenzelfde SCHAAL te vergelijken. Daarom kunnen twee leerders op een B1-niveau ingeschaald worden op basis van verschillende sterktes en zwaktes. Als het belangrijk is om een onderscheid te maken tussen vaardigheden in verschillende domeinen, dan zouden vaardigheden apart kunnen worden getoetst. Dan zouden er specifieke descriptoren kunnen worden gebruikt voor elke vaardigheid apart, als basis voor de beschrijving van vaardigheidsniveaus van specifieke vaardigheden.

Toch moet er rekening gehouden worden met enkele beperkingen wanneer het ERK aan een bepaalde context wordt aangepast. Het ERK is namelijk enkel bedoeld om taalvaardigheid te beschrijven in functie van het model van taalgebruik beschreven in 1.1.2 van deze handleiding. Er zou dus geen poging ondernomen mogen worden om kennis of vaardigheden te relateren die niet in dit model zijn opgenomen.

## 1.2 Validiteit

### 1.2.1 Wat is validiteit?

Validiteit kan eenvoudig worden gedefinieerd: een toets is valide als we meten wat we moeten meten. Als onze toets bijvoorbeeld is bedoeld om communicatieve vaardigheden in het Italiaans te meten en kandidaten scoren systematisch hoger of lager afhankelijk van hun vaardigheden in het Italiaans, dan is onze toets valide. Deze eerder nauwe definitie is de laatste jaren uitgebreid. Recentere theoriën definiëren validiteit als de manier waarop toetsen *gebruikt* worden. Dat betekent dat validiteit nu eerder verwijst naar 'de mate waarin de theorie en de gegevens de interpretatie van de scores, zoals die blijken uit hun gebruik, ondersteunen' (AERA, APA, NCME, 1999).

Deze uitgebreidere definitie benadrukt de sociale IMPACT van toetsen en de nood aan toetsen die adequate informatie verschaffen om eventueel belangrijke beslissingen te nemen over individuele kandidaten. In die zin kunnen we ook niet spreken van een valide toets in de absolute zin. Validiteit hangt immers samen met de manier waarop toetsresultaten voor een specifiek doel gebruikt worden: het is de interpretatie van de betekenis van de toetsresultaten van een individuele kandidaat die valide kan zijn of niet.

Volgens Bachman (1990) betekent dit voor taal dat toetsen de conclusie zouden moeten ondersteunen die wordt getrokken voor een bepaald domein waarin de doeltaal wordt gebruikt. Als we een uitspraak willen doen over de validiteit van een toets, dan moeten we dus eerst bepalen wat een kandidaat moet kunnen in de doeltaal in een levensechte situatie. Pas dan kunnen we zeggen of de toets voldoende bewijst dat de kandidaat hiertoe in staat is. Het ERK stelt een manier voor om de vaardigheden in specifieke gebruiksdomeinen te bepalen. De illustratieve descriptoren die het bevat, vormen een vertrekpunt.

### 1.2.2 Validiteit en het ERK

Als we resultaten rapporteren in termen van het ERK, dan beweren we dat we in staat zijn om prestaties op een toets te interpreteren overeenkomstig met onze definitie van wat kandidaten op een bepaald ERK-niveau moeten kunnen. Valideren betekent dan dat we bewijzen dat wat we beweren ook juist is: dat de leerder die als een B1 werd gerapporteerd ook een B1 *is* volgens het bewijs dat we kunnen leveren.

Het soort bewijs dat nodig is, varieert naargelang de context van de toets. Het ERK-model van taalgebruik en taalleren zoals hierboven voorgesteld kan sociaal-cognitief genoemd worden: taal is zowel een eigengemaakte set van vaardigheden, als een publieke set van sociale gedragingen. Afhankelijk van de context, zal een taaltoets meer op het ene dan op het andere focussen en dat beïnvloedt het bewijs dat de toets valide is:

- als de klemtoon meer op het gebruik ligt, dan wordt de validiteit bewezen door de taal die effectief gebruikt wordt voor verschillende communicatieve doeleinden.
- als de focus eerder op de vaardigheden ligt, dan is de validiteit verbonden met de cognitieve vaardigheden, strategieën en taalkennis die de conclusie onderbouwen over potentiële vaardigheden voor taalgebruik.

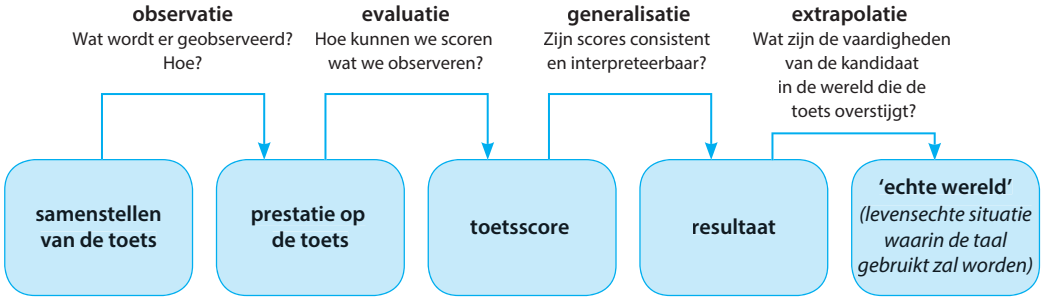
In het laatste geval zal het belangrijker zijn om aan te tonen dat de toetstaken dezelfde vaardigheden, strategieën en taalkennis vereisen die ook nodig zouden zijn in hetzelfde gebruiksdomein in de doeltaal, namelijk dat er *interactieve authenticiteit* is tussen beide (zie 1.1.3).

Beide soorten bewijs kunnen de ERK-gerelateerde validiteit van een taaltoets ondersteunen. Het evenwicht tussen beide hangt af van de eisen van een specifieke context. Een taaltoets voor verkopers zal waarschijnlijk veel belang hechten aan taalgebruik, terwijl een toets voor schoolkinderen meer gewicht zou kunnen geven aan vaardigheid.

### 1.2.3 Validiteit in het toetsontwikkelingsproces

Validiteit verbindt dus de prestatie op een toets met een conclusie over de taalvaardigheid van een kandidaat in een wereld die de toets overstijgt. Een toets ontwerpen en samenstellen is dus een essentiële stap, maar ook andere stappen zijn cruciaal.

In dit hoofdstuk wordt validiteit gerelateerd aan het ontwikkelingsproces van een toets (zie 1.5) zodat het duidelijk wordt wat de impact van elke stap is. Want, als de laatste fase valide moet zijn, moet elke voorafgaande fase van het ontwikkelingsproces goed doorlopen worden en een bevredigend resultaat opleveren.




**Figuur 3** ***De verschillende stappen die gezet worden als we denken over validiteit*** **(overgenomen van Kane, Crooks en Cohen, 1999; Bachman, 2005)**

Figuur 3 illustreert deze stappen op een schematische manier:

1. De toets is ontworpen om een aantal vaardigheden te ontlokken die interpreteerbaar zijn en die gebaseerd zijn op een model van vaardigheden van leerders. Een kandidaat kan bijvoorbeeld gevraagd worden om een brief te schrijven naar een vriend over een bepaald onderwerp.
2. Er wordt een score toegekend aan de productie van gesproken en geschreven taal van de kandidaat. Welke elementen van de responsen worden er beloond of afgestraft? In ons voorbeeld worden deze elementen gerelateerd aan de communicatieve vaardigheid beschreven in het model van taalgebruik. Dit model bevat REGISTER (sociolinguïstische vaardigheid), lexicon, grammatica en orthografische vaardigheid (linguïstische vaardigheden), enz.
3. Tot hier zijn de scores cijfers die samenhangen met een enkele productie van gesproken of geschreven taal voor een specifieke taak. Hoe kunnen die worden veralgemeend? Zou de kandidaat eenzelfde resultaat krijgen op een ander moment, voor een andere versie van de toets? Deze vragen hebben te maken met betrouwbaarheid (zie hoofdstuk 1.3). Een tweede aspect van generalisatie hangt samen met het afstemmen op een ruimere vaardigheidsschaal, omdat één vorm van een toets bijvoorbeeld makkelijker kan zijn dan een andere en we dit zouden willen onderkennen (zie bijlage VII).
4. Tot dusver hebben we de productie van gesproken en geschreven taal beschreven in de context van de toets, maar we willen die ook extrapoleren naar de wereld die de toets overstijgt. Hier relateren we een resultaat aan een ERK-niveau door te beschrijven wat een kandidaat zou moeten kunnen in de echte wereld. Daarvoor gebruiken we de 'can do'-descriptoren als gids.
5. We baseren ons hierop om beslissingen te nemen over de kandidaat.

Uit dit beknopte overzicht is het duidelijk dat validiteit, evenals de aanspraak die er gemaakt wordt op een link met het ERK, afhangen van elke stap in het ontwikkelingsproces en het afnemen van de toets. Validiteit is essentieel doorheen het hele proces van toetsontwikkeling.

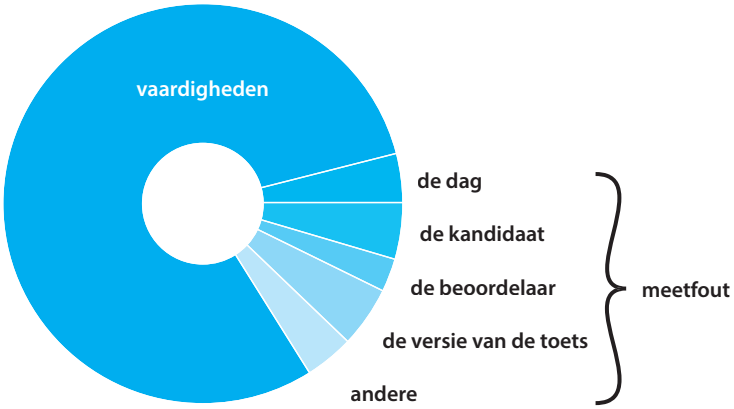
Bijlage I geeft richtlijnen over het opbouwen van een VALIDITEITSBEWIJS.

# 1.3 Betrouwbaarheid

## 1.3.1 Wat is betrouwbaarheid?

In een toetscontext betekent betrouwbaarheid consistentie: een toets met betrouwbare scores geeft keer op keer eenzelfde of een vergelijkbaar resultaat. Dit betekent dat kandidaten die eenzelfde toets afleggen altijd gerangschikt worden op bijna dezelfde manier. Dat betekent *niet* dat dezelfde mensen zouden slagen of zakken, want er kan een andere cesuur gehanteerd worden.

Merk op dat een hoge betrouwbaarheid niet noodzakelijk impliceert dat een toets goed is of dat de interpretatie van de resultaten valide is. Een slechte toets kan heel betrouwbare scores produceren. Het tegenovergestelde is echter niet waar: voor een valide interpretatie van toetsresultaten, *moeten* scores een acceptabele betrouwbaarheid hebben, want onbetrouwbare resultaten kunnen nooit betekenisvol zijn.




**Figuur 4** ***Een aantal bronnen van meetfout in een toetsscore***

Toetsscores variëren naargelang de kandidaten. De betrouwbaarheid wordt gedefinieerd als die proportie van de variatie in de toetsscore die ontstaat door de gemeten vaardigheden. De variatie die ontstaat door andere factoren wordt *MEETFOUT*[2] genoemd. Alle toetsen bevatten een bepaalde hoeveelheid meetfout.

Figuur 4 toont enkele veel voorkomende bronnen van meetfout:

- de dag waarop de toets wordt afgenomen (het weer, de afname, enz. kan anders zijn);
- de kandidaat zelf kan op een bepaalde dag beter of slechter presteren;
- de BEOORDELAARS die de versie van de toets beoordelen kunnen anders te werk gaan.

Er kunnen nog andere factoren zijn die we niet kunnen controleren.

We beogen toetsen te ontwikkelen waarbij verschillen in de scores vooral veroorzaakt worden door de verschillende vaardigheden van de kandidaten en niet door de meetfout.

## 1.3.2 Betrouwbaarheid in de praktijk

Toetsontwikkelaars zouden zich bewust moeten zijn van de bronnen van meetfouten, en ze zouden er alles aan moeten doen om de impact ervan tot een minimum te beperken. De procedures en concepten die in deze handleiding worden toegelicht, kunnen hierbij helpen. Het is ook belangrijk om de resultaten in statistische programma's in te voeren om op deze manier de betrouwbaarheid van de toetsscores te bepalen. Bijlage VII bevat meer informatie over hoe de betrouwbaarheid kan worden ingeschat.

Er kan geen betrouwbaarheidsgraad worden vooropgesteld die voor alle toetsen geldt, want de betrouwbaarheid hangt af van de mate waarin de toetsscores variëren. Een toets die ontwikkeld werd voor een beperkte groep van leerders die op basis van een aantal criteria werden geselecteerd, zal sowieso lager scoren op betrouwbaarheid omdat de groep kandidaten te klein en niet gevarieerd genoeg is. Betrouwbaarheid kan

2 Noot van de vertaalster: ook wel CONSTRUCT-IRRELEVANTE VARIANTIE genoemd.

bovendien afhangen van de items of het soort taken en van de manier waarop er beoordeeld wordt. De scores van taken die beoordeeld worden (zie hoofdstuk 5) zijn doorgaans minder betrouwbaar dan die van taken met DICHOTOME ITEMS waarbij scores worden toegekend door foute antwoorden te turven.

Toch is het zinvol om regelmatig de betrouwbaarheid te bestuderen, want zo kan er vastgesteld worden welke toetsen beter of slechter waren en kan er over de jaren heen geobserveerd worden hoe de kwaliteit van toetsen verbetert. De meeste indicatoren van betrouwbaarheid, zoals Cronbach's Alpha en KR-20, worden uitgedrukt in een cijfer tussen 0 en 1. Doorgaans wordt een betrouwbaarheidsgraad die in het bovenste derde deel van het bereik ligt (tussen 0.6 en 1) als aanvaardbaar beschouwd[3].

De betrouwbaarheid op een statistische wijze bepalen, is meestal niet mogelijk in situaties waar het aantal kandidaten en/of items laag is. In deze gevallen is het niet mogelijk om te bepalen of de betrouwbaarheid adequaat is voor de toets of niet. Een goede evaluatiestrategie in dit soort situaties is ervan uitgaan dat de toets slechts een deel van het bewijs levert dat nodig is om een beslissing te nemen. Aanvullend bewijs zou dan van een portfolio kunnen komen of van meerdere toetsen die verspreid over een bepaalde periode werden afgelegd.

## 1.4 Ethiek en rechtvaardigheid

### 1.4.1 Sociale gevolgen van toetsen: ethiek en rechtvaardigheid

Messick (1989) ijverde voor de rol van waarden en gevolgen van een toets als deel van de validiteit van een toets. Zijn invloed heeft ertoe geleid dat er meer aandacht werd besteed aan de sociale waarde van toetsen en de gevolgen voor de BELANGHEBBENDEN. De effecten en consequenties van toetsen bevatten de (hopelijk positieve) bedoelde uitkomsten van evaluatie alsook de onvoorziene en soms negatieve neveneffecten die toetsen kunnen hebben. De introductie van een nieuwe toets kan er bijvoorbeeld voor zorgen dat leerkrachten hun manier van lesgeven (positief of negatief) veranderen ('washback').

Toetsaanbieders kunnen onderzoek doen naar de washback en toetsimpact om meer te leren over de sociale gevolgen van hun toets. Dit soort onderzoek kan op heel kleine schaal uitgevoerd worden. In de klascontext is het mogelijk om te zien of studenten sommige aspecten van de syllabus als prioritair beginnen te beschouwen ten opzichte van andere delen, en of dit misschien komt door de focus van de toets. Om ze te motiveren om aan die aspecten te werken die wat verwaarloosd werden, kan bv. de focus van de toets verschuiven.

### 1.4.2 Rechtvaardigheid

Iedereen die toetsen aanbiedt, streeft ernaar een zo rechtvaardig mogelijke toets te ontwikkelen. Zie de *Code of Fair Testing Practices in Education* (JCTP, 1988) en de *Standards for Educational and Psychological Testing* (AERA et al., 1999).

De *Standards* uit 1999 erkennen drie aspecten van rechtvaardigheid: rechtvaardigheid in de vorm van het ontbreken van bias (zie Bijlage VII), rechtvaardigheid in de vorm van een onpartijdige behandeling tijdens het toetsproces en rechtvaardigheid in de vorm van gelijkheid in de uitkomsten van toetsing.

Kunnans *Test Fariness Framework* (Kunnan, 2000a, 2000b, 2004, 2008) focust op vijf aspecten van de evaluatie van taal waaraan voldaan moet worden om rechtvaardigheid te kunnen bereiken: *validiteit* (zie hoofdstuk 1.2), *afwezigheid van bias* (zie Bijlage VII), *toegang, afname* (zie hoofdstuk 4) en *sociale gevolgen*.

Verschillende instanties hebben *Codes of Practice* en *Codes of Fairness* ontworpen om toetsaanbieders hulp te bieden bij de praktische aspecten van het garanderen van rechtvaardige toetsen.

Toetsaanbieders kunnen proberen om bias te minimaliseren wanneer ze toetsen ontwerpen. Bepaalde onderwerpen (bv. lokale gebruiken) kunnen bijvoorbeeld bepaalde groepen kandidaten bevoordelen of benadelen (bv. diegenen die afkomstig zijn uit landen met heel andere tradities). Er kan aan de itemschrijvers een lijst worden gegeven met onderwerpen die vermeden moeten worden in toetsitems. Groepen die bepaald worden door de leeftijd, nationaliteit of het geslacht van de kandidaten kunnen belangrijk zijn, maar dit hangt af van de context van de toets (zie 3.4).

3 Noot van de vertaalster: Er bestaat geen gouden standaard voor betrouwbaarheid. De betrouwbaarheid van een toets is afhankelijk van factoren zoals het aantal items en bij de evaluatie van de betrouwbaarheid dient er rekening gehouden te worden met de consequenties die aan de toets verbonden zullen worden (high-stakes toetsen vereisen een hogere betrouwbaarheid).

### 1.4.3 Ethische bekommernissen

Sinds de vroege jaren 80 van de vorige eeuw worden ethische bekommernissen ook in taaltoetsing besproken. Spolsky (1981) bijvoorbeeld waarschuwde voor de negatieve gevolgen die HIGH-STAKES taaltoetsen op mensen kunnen hebben. Hij wees erop dat taaltoetsen gelabeld zouden moeten worden zoals medicijnen: 'gebruik met mate'. Hij focust vooral op het specifiek gebruik van taaltoetsen, bv. in een migratiecontext, waar de beslissingen op basis van een toetsscore grote en ernstige gevolgen kunnen hebben voor een persoon.

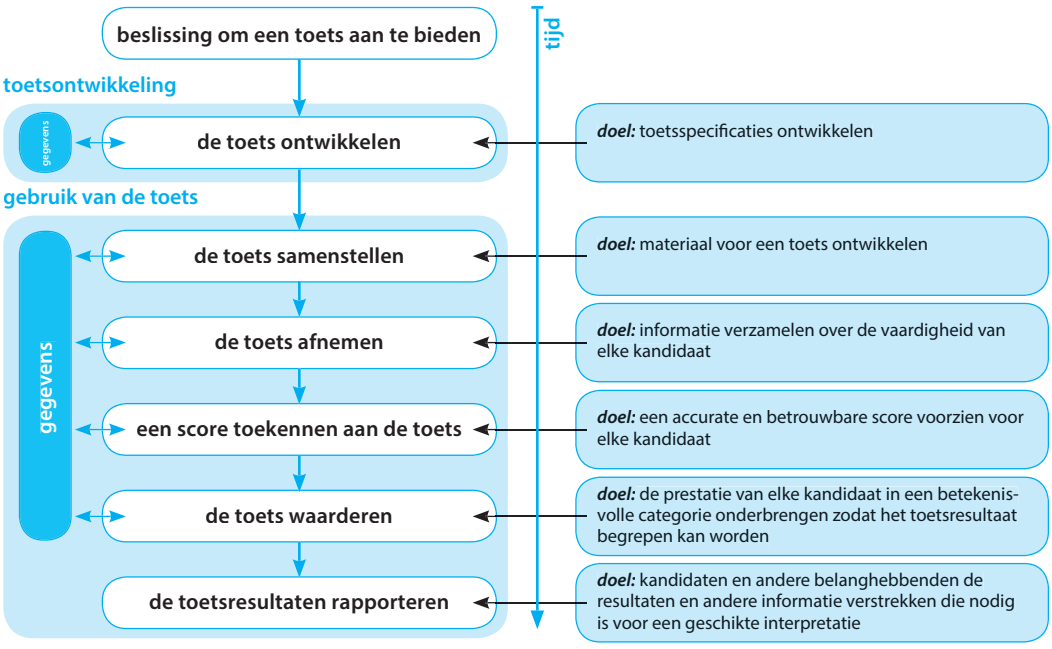
De *International Language Testing Association* (ILTA) publiceerde haar Code of Ethics in 2000. Daarin geeft ze richtlijnen over hoe toetsaanbieders zich professioneel zouden moeten gedragen. Toetsaanbieders moeten ervoor zorgen dat de relevante principes wijdverspreid en begrepen worden in alle geledingen van hun organisatie. Dit helpt om ervoor te zorgen dat de toetsinstelling deze richtlijnen ook in de praktijk brengt. Verdere maatregelen kunnen ook gepast zijn voor sommige aspecten van rechtvaardigheid in toetsen (zie hoofdstuk 4 en Bijlage VII).

## 1.5 Het werk plannen

De fase van de toetsontwikkeling en het gebruik van de toets vormen een cyclus waarin succes in een bepaald stadium afhangt van de uitkomsten van een voorgaande fase. Het is daarom belangrijk dat de hele cyclus goed beheerd wordt. Het is ook belangrijk om gegevens te verzamelen. Deze gegevens kunnen dan gebruikt worden om belangrijke beslissingen te maken tijdens het proces.

### 1.5.1 De fasen in het proces

Figuur 5 illustreert de stadia die doorlopen worden bij de ontwikkeling van een nieuwe toets. Het start met de beslissing om een toets aan te bieden. Die beslissing kan worden genomen door de aanbieder of door iemand anders, zoals een bestuur of ministerie. Daarna komt de fase van de toetsontwikkeling, gevolgd door die stadia die verbonden zijn met het gebruik van de toets. Elke fase bestaat uit kleinere stappen die allemaal moeten worden gezet om een fase als afgewerkt te kunnen beschouwen. Alle stappen samen zijn ontworpen om de doelen te bereiken die in de lijst aan de rechterkant van het diagram staan. Een tijdlijn toont aan dat deze stadia elkaar opvolgen, want de uitkomsten van een fase zijn nodig om een volgende fase te kunnen beginnen. Eenmaal de toets is ontwikkeld, kan de fase van het gebruik van de toets meerdere keren herhaald worden door de output van de toetsonwikkelingsfase (de toetsspecificaties) te hergebruiken. Op die manier kunnen meerdere EQUIVALENTE VORMEN van eenzelfde toets worden samengesteld.




**Figuur 5** ***De elementaire toetscyclus***

De stadia in Figuur 5 zijn toepasbaar op alle projecten van het samenstellen van een toets, hoe groot of hoe klein de toetsaanbieder ook is.

Elk van de fases vertegenwoordigd in Figuur 5 bevat vele kleinere 'micro'-taken en activiteiten. Deze worden meer in detail beschreven in de andere hoofdstukken van deze handleiding. De processen om een taak te volbrengen zouden gestandaardiseerd moeten zijn om ervoor te zorgen dat, elke keer er een versie van een toets samengesteld wordt, die ook vergelijkbaar is met de vorige versies.

Het verzamelen en gebruiken van gegevens wordt ook getoond aan de linkerkant van het diagram. Informatie over de achtergrond van de kandidaten, feedback van betrokkenen, of prestaties van kandidaten op taken en items en de tijd die ze nodig hebben om een taak uit te voeren, is belangrijk als een voortdurende controle dat de ontwikkeling naar wens verloopt en kan later ook helpen bij het aantonen dat het aanbevolen gebruik van de toetsresultaten valide is.

Plan de verzameling en het gebruik van dit soort gegevens in als een routineactiviteit. Anders zou dit wel eens vergeten kunnen worden tijdens het toetsontwikkelingsproces.

## 1.6 Kernvragen

- Welke aspecten van het ERK-model voor taalgebruik zijn het meest geschikt voor uw context?
- Welke ERK-vaardigheidsniveaus zijn het meest geschikt voor uw context?
- Hoe wilt u dat uw toetsresultaten worden begrepen en geïnterpreteerd?
- Wat is in uw context het belangrijkste dat de betrouwbaarheid in het gedrang kan brengen?
- Hoe kunt u er helpen voor zorgen dat uw werk zowel ethisch als rechtvaardig is voor de kandidaten?
- Met welke uitdagingen zult u geconfronteerd worden wanneer u de toetscyclus plant?

## 1.7 Aanbevolen literatuur

**Modellen van taalgebruik**

Fulcher en Davidson (2007, p. 36-51) bespreken nog meer constructen en modellen.

**Validiteit**

ALTE (2005, p. 19) geeft een nuttig overzicht van soorten validiteit en biedt achtergrond bij de moderne concepten van validiteit.

Kane (2004, 2006), Mislevy, Steinberg en Almond (2003) bespreken onderwerpen die gerelateerd zijn aan validiteitsbewijzen (Bijlage I van deze handleiding) en geven meer details over hoe ze ontwikkeld kunnen worden.

**Betrouwbaardheid**

Traub en Rowley (1991) en Frisbie (1988) beschrijven beiden op een toegankelijke manier de betrouwbaarheid van toetsresultaten. Parkes (2007) illustreert wanneer en hoe informatie van een bepaalde toets aangevuld kan worden met andere gegevens om beslissingen te nemen over kandidaten.

**Ethiek en rechtvaardigheid**

Sinds de vroege jaren 90 werden er deskundige *Codes of Practice* voor taaltoetsers ontwikkeld door verschillende professionele taaltoetsverenigingen zoals de ALTE *Code of Practice* (1994), de ILTA *Guidelines for Practice* (2007) of de EALTA *Guidelines for Good Practice in Language Testing and Assessment* (2006).

In de jaren 90 werkte Alan Davies (1997) mee aan een editie van *Language Testing* die focuste op de ethiek in taaltoetsing en in 2002 werd in Pasadena de *Language Assessment Ethics Conference* georganiseerd. Wat er op deze conference werd besproken, leidde tot een speciale uitgave van *Language Assessment Quarterly* (waaraan Davis ook meewerkte, 2004). McNamara en Roever (2006) bespreken overzichtsartikelen over rechtvaardigheid en ethische codes voor toetsing.

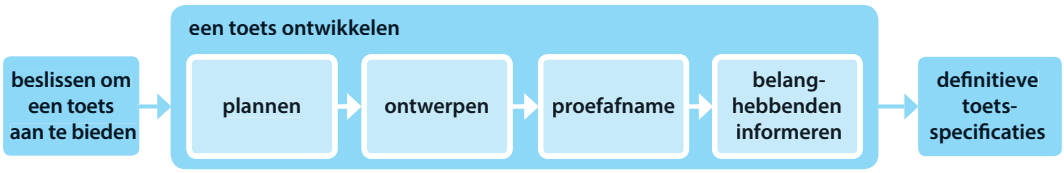
Gegevens verzamelen in functie van de rechtvaardigheid van een toets en dit in de vorm van een validiteitsarguments poneren, is het onderwerp van verschillende artikels in *Language Testing*, april 2010 (Davies, 2010; Xi, 2010).

# 2 Een toets ontwikkelen

## 2.1 Het toetsontwikkelingsproces

Tijdens het toetsontwikkelingsproces is het de bedoeling TOETSSPECIFICATIES vast te leggen voor de opbouw van de LIVE TOETS. De toetsontwikkeling begint met een persoon of organisatie (de sponsor van de toets) die beslist dat er nood is aan een nieuwe toets. Figuur 6 illustreert het toetsontwikkelingsproces en toont drie essentiële fases (plannen, ontwerpen en uitproberen) en een fase (BELANGHEBBENDEN informeren) die in sommige contexten nodig kan zijn. De verspreiding helpt namelijk niet bij het opstellen van de toetsspecificaties, maar heeft tot doel anderen in te lichten over de nieuwe toets.

Een gedetailleerder diagram is opgenomen in Bijlage II.




**Figuur 6** ***Het toetsontwikkelingsproces***

## 2.2 De beslissing om een toets aan te bieden

De beslissing om een toets aan te bieden wordt in deze handleiding niet gezien als een onderdeel van het toetsontwikkelingsproces. Toch geeft het belangrijke input voor de planningsfase. De eisen die door de opdrachtgever worden geïdentificeerd, zullen immers het ontwerp bepalen en de manier waarop de toets wordt gebruikt.

Wie beslist er of er een nieuwe taaltoets nodig is? In sommige gevallen wordt deze beslissing genomen door de toetsaanbieder die het toetsontwikkelingsproject ook zal uitvoeren. Soms hebben derden al beslist dat er nood is aan een toets en wordt de toetsaanbieder gevraagd om een toets te ontwikkelen.

In beide gevallen moeten de eisen duidelijk zijn, en dit kan extra werk betekenen voor wie de toets ontwikkelt. Het kan moeilijker zijn om de intenties van de sponsor te begrijpen omdat het niet om dezelfde organisatie gaat, of omdat de sponsor geen deskundige is op het gebied van taaltoetsing of –onderwijs. Soms weet hij ook niet welke informatie de toetsontwikkelaar nodig heeft.

## 2.3 De planning

Tijdens deze fase wordt er informatie verzameld die in latere fases zal worden gebruikt. De informatie zou vooral door de sponsor moeten worden verstrekt. Toch kunnen er ook andere belanghebbenden geconsulteerd worden zoals ministeries en beleidsorganen, uitgeverijen, scholen, ouders, deskundigen, werkgevers, onderwijsinstellingen en administratieve diensten. Worden er veel mensen bevraagd, dan kan dit proces gestructureerd worden aan de hand van vragenlijsten of werkgroepen. In een klascontext is persoonlijke kennis van de context en de kandidaten misschien voldoende. Dit zijn de belangrijkste vragen waarop een toetsontwikkelaar antwoord moet krijgen:

- Wat zijn de kenmerken van de kandidaten die de toets zullen afleggen? (leeftijd, geslacht, sociale situatie, opleiding, moedertaal, enz.)
- Waarvoor dient de toets? (diploma bij het verlaten van de school, toelating tot een opleiding, minimale professionele eisen, formatieve of diagnostische functie, enz. )
- Wat is de link met een onderwijscontext? (curriculum, methodologische benadering, leerdoelen, enz.)

- Welke standaard is er vereist voor het vooropgestelde doel? (een ERK-niveau voor bepaalde vaardigheden, een standaard die verbonden is met een bepaald domein, enz.)
- Hoe zullen de toetsresultaten worden gebruikt?

Beschikt de toetsontwikkelaar over het antwoord op deze vragen, dan kan hij beginnen met het definiëren van de taalvaardigheid die zal worden getoetst. Hij kan dan ook beslissen waar de cesuur moet liggen zodat er beslissingen kunnen worden genomen over kandidaten (zie hoofdstuk 5). Daarnaast kan hij beslissen hoe de resultaten gepresenteerd en uitgelegd kunnen worden aan de mensen die ze gebruiken (zie hoofdstuk 5).

Vragen over het effect van de toets in een bredere context zijn ook nuttig:

- Wie zijn de belanghebbenden?
- Welk soort impact is wenselijk?
- Welk soort impact wordt er verwacht?

Tot slot mogen ook de meer praktische vragen niet worden vergeten:

- Hoeveel kandidaten worden er verwacht?
- Tegen wanneer zou de toets klaar moeten zijn?
- Hoe zal de toets gefinancierd worden en welke budgetten kunnen er aangewend worden?
- Hoe vaak zal de toets worden afgenomen?
- Waar zal de toets worden afgenomen?
- Welke vorm van aanlevering is er vereist? (papieren of digitale versie)
- Wie zal er verantwoordelijk zijn voor elke fase van het aanbieden van de toets? (productie van het materiaal en het samenstellen van de toets, afname, beoordeling, communicatie van de resultaten)
- Welke gevolgen zal dit hebben voor de veiligheid van de toets? (bv. zou er meer dan één toets moeten worden gebruikt?)
- Hoe zal de toets worden gemonitord op langere termijn?
- Zal het mogelijk of praktisch haalbaar zijn om een pretest te organiseren? (zie hoofdstuk 3.4)
- Welke gevolgen zal dit hebben voor de logistiek? (zal het werk van de toetsaanbieder afhangen van andere organisaties zoals toetsinstellingen?)

## 2.4 Het ontwerp

De informatie uit de planningsfase wordt als vertrekpunt genomen voor de ontwerpfase. Er worden belangrijke beslissingen genomen over de eigenheid van de toets en er ontstaat een eerste versie van de toetsspecificaties. Deze beschrijven de algemene structuur van de toets en alle inhoudelijke aspecten. Meer gedetailleerde toetsspecificaties, bijvoorbeeld voor materiaalschrijvers en personeel dat betrokken is bij de afname en het organiseren van de toets, kunnen ontwikkeld worden nadat deze eerste specificaties goedgekeurd werden.

### 2.4.1 Aandachtspunten voor de beginfase

De eerste uitdaging in de ontwerpfase is de ontwikkeling van een duidelijker idee van de inhoud en het format van de toets. Dit begint met de toetseisen en achtergrondinformatie, zoals de kenmerken van de kandidaten, het doel van de toets en het vereiste vaardigheidsniveau.

Het ERK is een nuttige bron om de eigenschappen van de toets te definiëren omdat het veel relevante hoofdstukken over toetsing bevat, meer bepaald:

- hoofdstuk 6, over het leren en onderwijzen van talen, nodigt uit om na te denken over de leerdoelen en de onderwijsmethodologie. Dit heeft dan weer een impact heeft op de stijl, de inhoud en de functie van een toets.

- hoofdstuk 7, over TAKEN en hun rol in het taalonderwijs, beschrijft hoe taken in toetsing kunnen worden gebruikt.
- hoofdstuk 9, over evaluatie, bespreekt hoe het ERK voor verschillende doeleinden kan worden gebruikt.

Het ligt voor de hand dat vooral hoofdstuk 4 en 5 relevant zijn. Daarin wordt de inhoud van de toets beschreven en de vaardigheden die moeten worden getoetst. De ontwerper van de toets kan selecteren uit de vele opties die de actiegerichte aanpak en het model voor taalgebruik van het ERK bieden (zie hoofdstuk 1.1) zoals:

- de focus van taken, bv. begrip tonen van details van een tekst, enz. – zie ERK (hfdst. 4.4 en 4.5)
- wat moet er getoetst worden, bv. vaardigheden, strategieën – zie ERK (hfdst. 5)
- de tekstsoorten die als INPUT worden gebruikt – zie ERK (hfdst.4.6)
- tekstbronnen – zie ERK (hfdst. 4.1 en 4.6)
- een indicatie van welke onderwerpen als geschikt worden beschouwd – zie ERK (hfdst. 4.1 en 4.2)
- het soort RESPONSSTIMULUS dat in mondelinge toetsen wordt gebruikt – zie ERK (hfdst.4.3 en 4.4)
- soorten levensechte situaties die relevant zijn voor de kandidaten – zie ERK (hfdst. 4.1 en 4.3)
- het niveau van prestatie dat nodig is in dit soort situaties - zie de vele illustratieve 'Can do'-SCHALEN in het ERK
- criteria om schriftelijke OPEN TAKEN en mondelinge toetsen te evalueren – zie relevante illustratieve 'Can do'-schalen in het ERK, bv. p. 58 en p. 74 enz.

De toetsaanbieder moet ook een aantal technische kenmerken van de toets vastleggen zoals:

- De duur van de toets. Er zou genoeg tijd moeten zijn zodat een doorsnee kandidaat voldoende tijd heeft om de toets af te werken zonder dat hij zich moet haasten. Het belangrijkste is dat kandidaten voldoende mogelijkheden hebben om hun werkelijke vaardigheden te tonen. Dit moet misschien eerst worden ingeschat door een ervaren taaltoetser. Ook voorbeelden uit de literatuur kunnen verhelderend werken (zie Aanbevolen literatuur, hoofdstuk 2.8). Eenmaal er een PROEFAFNAME is gebeurd, of de toets in operationeel gebruik is, kan de tijdsduur gecontroleerd worden. Soms worden er toetsen gebruikt waarbij een zekere snelheid wordt verwacht en kandidaten worden aangemoedigd om snel te antwoorden. Ook in dit geval moet de toegekende tijd geëvalueerd worden.
- Het aantal items in de toets. Er zijn voldoende items nodig om de toets van de nodige inhoud te voorzien en voldoende betrouwbare informatie over de vaardigheid van de kandidaat te kunnen verzamelen. Toch zijn er, omwille van de praktische haalbaarheid, ook grenzen aan de lengte van de toets.
- Het aantal items per onderdeel. Als de toets bedoeld is om veel verschillende vaardigheden te meten, dan zullen er voldoende items per onderdeel nodig zijn. Er kunnen voorbeelden geconsulteerd worden en de betrouwbaarheid kan berekend worden (zie Bijlage VII).
- De soorten items. Een item kan een gesloten of open respons ontlokken. Gesloten types zijn meerkeuze, matching en rangschikken. Open types omvatten korte antwoorden (tekst met gaten) of een uitgebreide schriftelijke respons. Elke soort items heeft voor- en nadelen. Zie ALTE (2005, p. 111-34) voor meer informatie over soorten items.
- De totale en individuele lengte van teksten, bv. in aantal woorden uitgedrukt. Voorbeelden uit de literatuur (zie Aanbevolen literatuur, hoofdstuk 2.8) kunnen richtlijnen geven over de praktisch haalbare lengte.
- Het format van een toets. Een toets met DISCRETE ITEMS bestaat uit korte items die onderling niet met elkaar verbonden zijn. In een taakgerichte toets worden toetsitems gebundeld in een kleiner aantal toetstaken, bv. gerelateerd aan een lees- of luistertekst. Omdat taakgerichte toetsen gebruik maken van langere, meer authentieke stimuli zijn ze vaak meer geschikt voor communicatieve taaltoetsing. Zie ALTE (2005, p. 135-47) voor meer informatie over soorten taken.
- De score die aan elk item wordt toegekend en de totaalscore voor elke taak of TOETSONDERDEEL. Hoe hoger de score voor elk item of onderdeel, hoe groter het relatieve belang. Meestal is het beter om per item eenzelfde score te voorzien. Toch kan er soms een reden zijn om items te WEGEN door ze een relatief grotere of kleinere proportie van het totaal aantal punten toe te kennen (zie Bijlage VII).

- De kenmerken van de BEOORDELINGSSCHALEN. Zullen er taakspecifieke schalen worden gebruikt, hoe uitgebreid zal elke schaal zijn, zullen de schalen analytisch of holistisch zijn? Hoofdstukken 2.5 en 5.1.3 behandelen beoordelingsschalen meer in detail.

Op het einde van de ontwerpfase zullen dus de eerste beslissingen worden genomen over het doel van de toets, de vaardigheden en onderwerpen die zullen worden getoetst en de details over de technische implementatie. We zouden ook moeten overwegen hoe de taken gecorrigeerd moeten worden, hoe beoordelingsschalen voor productieve vaardigheden (bv. voor schrijven en spreken) moeten worden ontwikkeld (zie hoofdstuk 2.5), hoe toetsen afgenomen moeten worden (hoofdstuk 4) en hoe CORRECTOREN en BEOORDELAARS getraind en opgevolgd moeten worden (hoofdstuk 5.1.3). Al deze beslissingen zouden dan door de belanghebben moeten worden geëvalueerd.

Communicatie met kandidaten en andere belanghebbenden kan betrekking hebben op de volgende zaken:

- Het geschatte aantal uren studie dat nodig is als voorbereiding voor de toets (Als er een formele studie wordt verwacht.);
- Hoe voorbeeldtoetsen ter beschikking zullen worden gesteld;
- Welke informatie er aan de gebruikers (alle relevante belanghebbenden) van de toets wordt gegeven, zowel voor als na de toets.

Tot slot moeten de verwachtingen van de belanghebbenden in acht worden genomen:

- Hoe zal de toets in het huidige systeem passen in termen van curriculumdoelen en klaspraktijk?
- Hoe verwachten de belanghebbenden dat de toets er zal uitzien?

Vooral hoofdstuk 4 van het ERK geeft een nuttig referentieschema om onderscheidende kenmerken van om het even welke toets in het ontwikkelingsproces duidelijker in kaart te brengen. Om dit te kunnen doen, wordt er een samenvatting van de toets weergegeven in de vorm van een schema. In Bijlage III van deze handleiding wordt deze methode verder toegelicht. De voorbeeldtoets bestaat uit vier onderdelen en is bedoeld voor leerders op B2-niveau die de taal studeren in een zakelijke context. Er wordt een overzicht gegeven van de inhoud van de hele toets en een algemene beschrijving voor een taak uit de toets.

### 2.4.2 De eisen versus de praktische haalbaarheid

In deze fase van de toetsontwikkeling moeten de praktische beperkingen van het voorgestelde toetsontwerp van naderbij bekeken worden. Informatie hierover wordt in de planningsfase verzameld, wanneer de eisen voor de toets worden vastgelegd (zie hoofdstuk 2.3). De toetsontwikkelaar moet de eisen met de beperkingen verzoenen en de sponsor van de toets moet het hiermee eens zijn. Bachman en Palmer (1996, hfdst. 2) geven een referentiekader om dit te doen aan de hand van hun concept van het NUT VAN EEN TOETS. Volgens hen is het NUT VAN EEN TOETS een functie van zes kwaliteiten:

- VALIDITEIT – de interpretaties van toetsresultaten of andere uitkomsten zijn betekenisvol en geschikt.
- BETROUWBAARHEID – de toetsresultaten die geproduceerd werden zijn consistent en stabiel.
- AUTHENTICITEIT – de taken lijken op levensechte taalactiviteiten in het/de betreffende domein(en).
- INTERACTIVITEIT – de taken zetten mentale processen en strategieën in gang die ook in levensechte situaties nodig zouden zijn.
- IMPACT – het effect, hopelijk positief, dat een toets heeft op individuen, de klaspraktijk en de bredere samenleving.
- PRAKTISCHE HAALBAARHEID – het zou mogelijk moeten zijn om een toets te ontwikkelen, samen te stellen en af te nemen zoals gepland, met de beschikbare middelen.

Deze kwaliteiten van een toets kunnen met elkaar concurreren: een verhoogde authenticiteit in de taken kan leiden tot een afname van de betrouwbaarheid bijvoorbeeld. Daarom is het belangrijk om te zoeken naar de beste balans zodat het algemene nut van de toets toeneemt.

### 2.4.3 Toetsspecificaties

De output van de toetsontwikkeling is een set van toetsspecificaties. De eerste kladversie van deze specificaties bevat beslissingen over veel van de informatie die werd besproken. Na de proefafname (zie hoofdstuk 2.5) zullen deze toetsspecificaties afgewerkt worden. Als het om een high-stakes toets gaat, zullen de toetsspecificaties een heel belangrijk hulpmiddel zijn om de kwaliteit van de toets te garanderen en anderen te tonen dat de aanbevolen interpretaties van de toetsresultaten valide zijn.

Toetsspecificaties zijn ook belangrijk voor low-stakes toetsen omdat ze er mee voor zorgen dat de verschillende toetsvormen dezelfde basis hebben, en dat de toets correct gerelateerd wordt aan de syllabus waarmee les wordt gegeven of aan andere kenmerken van de context waarin er wordt getoetst.

Toetsspecificaties kunnen op verschillende manieren beschreven worden, afhankelijk van de behoeften van de toetsaanbieder en het beoogde publiek. Er werden een aantal modellen van toetsspecificaties ontwikkeld (zie Aanbevolen literatuur, hoofdstuk 2.8) die als een zinvol vertrekpunt kunnen worden gebruikt.

## 2.5 De proefafname

Het is de bedoeling om in deze fase de kladversie van de toetsspecificaties uit te testen en verbeteringen aan te brengen, gebaseerd op de ervaringen met de proefafname en de suggesties van belanghebbenden.

Eenmaal de toetsspecificaties zijn uitgeschreven, worden er voorbeeldmaterialen ontwikkeld. Dit kan gedaan worden aan de hand van de richtlijnen die in hoofdstuk 3 van deze handleiding worden beschreven. Informatie over deze materialen kan op veel manieren worden verzameld:

- door een PILOTSTUDIE (een aantal kandidaten vragen om de toets af te leggen) te organiseren en de analyse van de responsen (zie hoofdstuk 3.4 en Bijlage VII);
- door collega's om advies te vragen;
- en door andere belanghebbenden om advies te vragen.

Een pilotstudie zou moeten worden uitgevoerd met kandidaten die representatief zijn voor de kandidaten waarvoor de toets wordt ontwikkeld. De pilot zou moeten worden afgenomen onder omstandigheden die lijken op een reële toetssituatie. Een pilot kan ook zijn nut hebben als de situatie van een live toets niet helemaal kan nagebootst worden (omdat er misschien te weinig tijd is om de volledige toets af te nemen, of om andere redenen), of als er maar weinig kandidaten beschikbaar zijn. De pilot kan informatie verschaffen over de tijd die nodig is voor een taak, de duidelijkheid van de INSTRUCTIES, de vormgeving voor de respons, enz. Voor mondelinge toetsonderdelen wordt een observatie van de respons (bv. in de vorm van een opname) aangeraden.

Collega's of belanghebbenden om advies vragen, kan op verschillende manieren. Gaat het om een kleine groep, dan is een face-to-facegesprek mogelijk. Voor grotere projecten kunnen er vragenlijsten of feedback in de vorm van een verslag gebruikt worden.

De informatie uit de pilot maakt het ook mogelijk om uitgebreide CORRECTIESCHEMA'S en beoordelingsschalen te ontwerpen (zie hoofdstuk 5.1.3 voor de kenmerken van beoordelingsschalen). De kenmerken kunnen geïdentificeerd worden aan de hand van prestaties van kandidaten die de vaardigheidsniveaus het best illustreren. Zij kunnen de basis vormen voor de descriptoren op elk niveau. Eenmaal ze opgebouwd zijn, worden de beoordelingsschalen in een pilot uitgetest. Daarbij wordt er zowel kwalitatief als kwantitatief geanalyseerd hoe de beoordelaars de beoordelingsschalen gebruiken (zie Bijlage VII). Daarna kan er nog een extra pilot nodig zijn, of kunnen er aanpassingen overwogen worden.

Andere vormen van onderzoek kunnen nodig zijn om een antwoord te bieden op vragen die tijdens de proefafname naar voren zijn gekomen. Het kan mogelijk zijn om de data uit de pilot hiervoor te gebruiken, maar misschien zijn er aanvullende onderzoeken nodig. Bijvoorbeeld:

- Zullen de taaktypes die we willen gebruiken goed werken bij de specifieke populatie waarvoor de toets is bedoeld (bv. kinderen)?
- Zullen de taaktypes valide zijn in het domein waarvoor de toets is bedoeld (bv. toerisme of recht)?

- Toetsen items en taken echt de vaardigheden die ze zouden moeten toetsen? Statistische technieken kunnen gebruikt worden om te bepalen hoe goed items en taken verschillende vaardigheden toetsen (zie Bijlage VII).
- Zullen beoordelaars in staat zijn om de beoordelingsschalen en de evaluatiecriteria te interpreteren en te gebruiken zoals ze bedoeld zijn?
- Als er een toets wordt aangepast, is er dan een vergelijkende studie nodig om te garanderen dat het nieuwe toetsontwerp op eenzelfde manier zal werken zoals de bestaande toets?
- Nodigen items en taken de kandidaten uit om mentale processen te gebruiken zoals beoogd? Dit zou kunnen worden bestudeerd via verbale protocollen, waarbij leerders hun denkprocessen verwoorden terwijl ze taken uitvoeren.

Toetsspecificaties kunnen verschillende keren gereviseerd worden vooraleer ze de vorm aannemen die zal worden gebruikt tijdens de live toets.

## 2.6 Belanghebbenden informeren

Toetsspecificaties worden op verschillende manieren gebruikt. Ze kunnen gelezen worden in functie van het schrijven van items, de voorbereiding van de toets en om te beslissen wat er moet worden onderwezen. In veel contexten zal dit betekenen dat er verschillende versies worden gemaakt voor verschillende gebruikers. Er zou bijvoorbeeld een vereenvoudigde versie kunnen worden ontwikkeld die het linguïstische bereik, onderwerpen, format enz. beschrijft voor de kandidaten die zich willen voorbereiden op de toets. Een meer gedetailleerd document kan worden gebruikt door de itemschrijvers.

Als aanvulling op de toetsspecificaties kunnen belanghebbenden het heel nuttig vinden om voorbeeldmaterialen te zien (zie hoofdstuk 3 voor meer informatie over materialen produceren). Als dit opportuun is, kunnen de voorbeeldmaterialen niet alleen de taken bevatten, maar ook audio- of video-opnames die in de luistertaken worden gebruikt. In een klascontext kunnen deze materialen voor een stuk als voorbereiding op de toets dienen. In de jaren die daarna volgen, kunnen gebruikte toetsversies als voorbeeldmateriaal worden aangewend.

Voor de kandidaten kan het zinvol zijn om antwoorden op voorbeelden van spreek- en schrijftaken te zien. Deze kunnen beschikbaar worden gemaakt als de materialen in het verleden werden uitgetest of werden gebruikt bij een live toets. Het is ook mogelijk om de kandidaten advies te geven dat hen helpt om zich voor te bereiden op de toets.

Het materiaal moet beschikbaar worden gesteld ruim voor de kandidaten het nodig hebben. Als er nog materiaal nodig zou zijn, zoals regelgeving, rollen, verantwoordelijkheden en uurroosters, dan zou dat ook op voorhand klaar moeten zijn.

## 2.7 Kernvragen

- Wie heeft er beslist dat er een toets moest komen? – Wat kunnen zij vertellen over het doel en het gebruik van de toets?
- Wat zal de educatieve en sociale impact zijn van de toets?
- Welk type en niveau van taalvaardigheid moet er getoetst worden?
- Welk soort taken is er nodig om dit te kunnen bereiken?
- Over welke praktische middelen beschikt u? (vb. ruimte, personeel, enz.)
- Wie zou er moeten worden betrokken bij het uitschrijven van de kladversie van de TOETSSPECIFICATIES en het ontwikkelen van voorbeeldtoetsmateriaal? (bv. in termen van deskundigheid, impact, autoriteit, enz.)
- Hoe zullen de inhoud, de technische en procedurele details van de toets beschreven worden in de toetsspecificaties?
- Welk soort informatie over de toets moet er aan de gebruikers worden gegeven? (bv. een publiekelijk beschikbare versie van de toetsspecificaties en hoe moet die verspreid worden)

- Hoe kan de toets uitgeprobeerd worden?
- Hoe kunnen belanghebbenden het best geïnformeerd worden over de toets?

## 2.8 Aanbevolen literatuur

Er bestaan een groot aantal ERK-gerelateerde voorbeelden van toetsmateriaal om diegenen die betrokken zijn bij het toetsontwikkelingsproces te helpen met het begrijpen van de ERK-niveaus. Zie Raad van Europa (2006 a,b; 2005), Eurocentres/Federation of Migros Cooperatives (2004), University of Cambridge ESOL Examinations (2004), CIEP/Eurocentres (2005), Bolton, Glaboniat, Lorenz, Perlmann, Balme en Steiner (2008), Grego Bolli (2008), Raad van Europa en CIEP (2009).

Voorbeeldversies van toetsspecificatieformats zijn te vinden in Bachman en Palmer (1996, p. 335-34), Alderson, Clapham en Wall (1995, p. 14-17) en Davidson en Lynch (2002, p. 20-32).

Er zijn een aantal roosters beschikbaar die templates voorzien voor het beschrijven en vergelijken van taken, onder andere ALTE Members (2005 a, b; 2007 a,b), Figueras, Kuijper, Tardieu, Nold en Takala (2005).

# 3 Een toets samenstellen

## 3.1 Het proces van een toets samenstellen

De bedoeling van deze fase is om op tijd materialen te voorzien voor de live toets, die in overeenstemming met de toetsspecificaties worden ontwikkeld. Het proces van een toets samenstellen verloopt in drie grote stadia, zoals in Figuur 7 wordt geïllustreerd.


**Figuur 7** ***Grote stadia in het proces van een toets samenstellen***

Het ontwikkelen van taken en het produceren van een toets worden hier beschreven als twee verschillende fases, omdat het makkelijker is om zo de doelstellingen van elke stap duidelijk te maken. Als materialen worden ontwikkeld en ze tegelijkertijd in de vorm van een toetsformulier worden gegoten, zijn de principes echter hetzelfde. Er moet daarna een kwaliteitscontrole volgen, met de mogelijkheid om de toetsmaterialen indien nodig aan te passen.

## 3.2 Voorbereidende stappen

Voor er materialen worden ontwikkeld, moet er over de volgende stappen worden nagedacht:

- de rekrutering en training van itemschrijvers;
- het beheer van materialen.

### 3.2.1 Itemschrijvers rekruteren en opleiden

Itemschrijvers kunnen dezelfde personen zijn die de toets ontwikkelen. In dit geval is rekrutering niet nodig en is de training relatief eenvoudig omdat de itemschrijvers al vertrouwd zijn met de toets en de doelstellingen die ermee nagestreefd worden.

Als er itemschrijvers moeten worden gezocht, moet de toetsaanbieder beslissen aan welke minimale professionele eisen zij moeten voldoen. Dit kan het niveau van taalvaardigheid omvatten, of het begrijpen van de context van de toets. Andere belangrijke aspecten voor het schrijven van items, zoals kennis van een bestaande toets of de evaluatieprincipes, kunnen in een training behandeld worden (zie ALTE, 2005). Die hoeven dus niet tot de minimale vereisten gerekend te worden. Een opleidingstraject, monitoring en evaluatie zullen er mee voor zorgen dat de itemschrijver zich professioneel verder ontplooit.

Taaldocenten kunnen goede itemschrijvers zijn omdat zij de taalleerders en de taal doorgaans goed begrijpen. Zij zijn nog geschikter als ze al kandidaten hebben voorbereid op een gelijkaardige toets of als ze al ervaring hebben met corrigeren of het afnemen van mondelinge toetsen. Itemschrijvers kunnen voor alle onderdelen van de toets ingezet worden, of enkel voor specifieke toetsonderdelen, afhankelijk van hun kennis en van de eisen van de toetsaanbieder.

### 3.2.2 Materiaal beheren

De toetsaanbieder moet een systeem uitwerken om items te verzamelen, op te slaan en te verwerken. Dit is vooral belangrijk als het over veel items en taken gaat. Alle toetsmaterialen moeten hetzelfde proces van kwaliteitscontrole doorlopen, zoals de vormgeving en de pilot. Daarom zou het mogelijk moeten zijn om op elk moment te zien in welke fase van het proces een item zich bevindt. Dit wordt nog belangrijker als er grote

aantallen items ontwikkeld worden en in de verschillende fase meer mensen betrokken zijn. Een elementair systeem van materiaalbeheer zou het volgende kunnen omvatten:

- Een uniek identificatienummer voor elk item;
- Een checklist waarmee vastgelegd wordt welke fases voltooid zijn, veranderingen werden doorgevoerd en andere informatie;
- Een manier om ervoor te zorgen dat items en gerelateerde informatie toegankelijk is en dat de versies van vroegere vormgevingsfases niet in circulatie zijn; misschien door ze te bewaren op één enkele locatie, of door alles formeel via e-mail door te geven op het einde van elke schrijf- of vormgevingsfase.

## 3.3 Materiaal produceren

Er wordt itemschrijvers gevraagd om materialen te produceren die gebruikt zullen worden in een live toets. In deze handleiding wordt aan deze stap gerefereerd als het 'ter beschikking stellen'. Bijlage IV bevat informatie die itemschrijvers kan helpen bij deze taak. Zij moeten weten hoeveel en welke soort items er nodig zijn en wanneer ze klaar moeten zijn.

Dit hoofdstuk focust op het bepalen van de materialen die vereist zijn en hoe dit aan de itemschrijvers wordt gecommuniceerd. Om te beslissen wanneer ze klaar moeten zijn, wordt er teruggeteld vanaf de datum waarop de live toets zal plaatsvinden.

### 3.3.1 Eisen voor de beoordeling

Om een toets te kunnen opbouwen, moeten toetsaanbieders een selectie aan taken en items hebben waaruit ze kunnen kiezen. Het is moeilijk om te bepalen hoeveel taken er precies ter beschikking moeten worden gesteld zodat de toets die wordt opgebouwd een evenwichtig aanbod aan kenmerken omvat wat betreft soort items, onderwerp, linguïstische focus en moeilijkheidsgraad (zie hoofdstuk 2.5). Dit betekent dat er meer items moeten worden ter beschikking gesteld dan er uiteindelijk zullen worden gebruikt. Een andere reden om meer items te voorzien, is dat sommige items hoogstwaarschijnlijk zullen worden verworpen tijdens de kwaliteitscontrole.

### 3.3.2 Het aanbieden van de toets

Materialen kunnen voor een welbepaalde afname ter beschikking worden gesteld of ze kunnen toegevoegd worden aan een ITEMBANK waaruit later zal geput worden om een andere toets samen te stellen. In beide gevallen moet er voldoende tijd voorzien worden voor alle stappen die moeten worden gezet bij het samenstellen van een toets.

Er zouden een aantal parameters moeten worden afgesproken om items te produceren. De belangrijkste doelstelling is om verwarring of misverstanden te voorkomen. Een langere lijst met formele eisen en vragen is zinvol als de groep itemschrijvers groot en divers is. Het is echter altijd nuttig om het eens te zijn over een aantal punten die hieronder worden beschreven, en om die punten ook vast te leggen:

**Details van de vereiste materialen**

Dit zal details bevatten:

- over het aantal teksten, taken en items dat is vereist;
- voor teksten: of de items onmiddellijk moeten worden geschreven of dat dit pas gebeurt nadat de tekst werd aanvaard;
- voor mondelinge toetsen met visuele responsstimulus: waar er een visuele responsstimulus wordt verwacht, of een indicatie van welke soort responsstimulus er nodig is;
- voor problemen die te maken hebben met copyright van afbeeldingen of teksten en hoe ermee moet worden omgegaan;

en vragen met betrekking tot:

- een SLEUTEL of correctieschema voor elk item, met de correcte respons:

- schriftelijke toetsen: voorbeeldantwoorden om ervoor te zorgen dat de taak met het maximaal vooropgestelde aantal woorden kan beëindigd worden en het taalvaardigheidsniveau van de potentiële kandidaten;
- een ingevuld formulier dat de taak op een gestandaardiseerde manier beschrijft.

**Details over hoe de materialen zouden moeten worden gepresenteerd**

- Het meest zinvolle format is een digitale kopie: deze kan makkelijk bewaard worden en de itemschrijvers kunnen in een bestaande template werken waardoor het format consistent blijft.
- Als er een volledige toets wordt geschreven, moet er nagedacht worden over de nummering en volgorde waarin de items voorkomen: worden ze achter elkaar genummerd en volgen de onderdelen elkaar op of komt elk onderdeel of elke oefening op een apart blad?
- Er moet ook overwogen worden hoe de bijdrages van de itemschrijvers worden gelabeld, bv. met de naam van de itemschrijver, datum en naam van de toets.

(Al deze details kunnen worden vastgelegd in de richtlijnen voor de itemsschrijvers – zie hieronder.)

**Details voor de deadline waarop de materialen moeten worden afgeleverd**

Er zou itemschrijvers moeten worden verteld wanneer hun materiaal zal worden vormgegeven en of zij worden verondersteld hierbij te helpen. Als itemschrijvers niet betrokken zijn bij de rest van het toetsontwikkelingsproces, kan hen verteld worden hoe hun rol past binnen de ruimere context van de ontwikkeling. Dit zal hen helpen om te begrijpen hoe belangrijk het is om zich aan de afgesproken deadlines te houden.

**Andere details zoals de arbeidsvoorwaarden**

Sommige itemschrijvers zullen de arbeidsvoorwaarden moeten kennen die voor hen gelden, zeker wanneer items schrijven bij ander werk binnen de organisatie komt kijken of wanneer er met freelancers wordt gewerkt. Er kan beslist worden om enkel te betalen voor bruikbare items (waardoor afgewezen materialen dus niet vergoed worden). Itemschrijvers kunnen ook worden betaald wanneer ze een eerste versie inleveren en kunnen bij de aanvaarding van de afgeleverde materialen een extra vergoeding krijgen. De verloning kan variëren naargelang het type item, of er kan een bepaalde som betaald worden voor een volledig onderdeel of een hele toets.

Leerkrachten die worden gevraagd om materiaal te schrijven voor toetsen die op een school zullen worden afgenomen, zouden voldoende tijd moeten krijgen om deze materialen tijdens hun uren te ontwikkelen.

Deze documenten zouden aan de itemschrijvers moeten worden gegeven:

- Gedetailleerde toetsspecificaties die geschikt zijn voor itemschrijvers. Dit kunnen vertrouwelijke documenten zijn, die meer details bevatten dan de publiek toegankelijke versie. Ze zouden gedetailleerd advies moeten bevatten over de presentatie van materialen. Zo verliezen itemschrijvers geen tijd met het maken van, waarschijnlijk foute, veronderstellingen over wat aanvaardbaar is.
- Voorbeeldmaterialen of toetsen die in het verleden werden gebruikt.

Er zou itemschrijvers ook moeten worden verteld wat het profiel van de kandidaten is, onder andere wat betreft hun leeftijd, geslacht en linguïstische achtergrond.

Afhankelijk van de context van de toets kan het nodig zijn om aanvullende richtlijnen te verstrekken zoals:

- een formulier dat de itemschrijver moet ondertekenen om te bevestigen dat hij de opdracht aanvaardt;
- een overeenkomst waarin staat dat de toesaanbieder zal beschikken over het copyright van de toetsmaterialen;
- een lijst of lexicon die het bereik en het niveau van de woordenschat en/of de structuren definieert die moeten worden gebruikt;
- een handboek dat informatie verstrekt over de toetsaanbieder.

## 3.4 Kwaliteitscontrole

### 3.4.1 Nieuw materiaal ontwikkelen

Als de toetsmaterialen zijn ingediend, moet de kwaliteit ervan worden nagegaan. Dit wordt gedaan aan de hand van het oordeel van deskundigen, maar ook door de nieuwe items uit te proberen. Als er een of meerdere dingen veranderen aan een item of aan een taak, dan zouden die opnieuw beoordeeld moeten worden op geschiktheid.

Dit soort controle is essentieel, maar elk item en elke taak zou, idealiter, niet door zijn auteur moeten worden nagekeken. In contexten met beperkte middelen, kunnen items en taken door een kleine groep collega's worden gecontroleerd. Als de itemschrijver alleen werkt, kan hij objectiever te werk gaan door extra tijd te laten tussen het produceren van de taken en de controle ervan, en door zoveel mogelijk items in één keer na te kijken.

Met de allereerste controle zou moeten worden nagegaan of de materialen beantwoorden aan de toetsspecificaties en aan alle andere afspraken die werden gemaakt. Er zou aan de itemschrijvers feedback moeten worden gegeven om ze toe te laten hun werk te controleren en hun vaardigheden als itemschrijvers verder te ontwikkelen. Er zou hen bijvoorbeeld kunnen gesuggereerd worden hoe een item kan worden veranderd (zie Bijlage V).

Teksten kunnen ook ingediend worden zonder items. Als de tekst aanvaard wordt, dan start de itemschrijver met het schrijven van de items voor de tweede fase van de vormgeving.

Deze eerste controle kan snel gebeuren en er kunnen relatief veel items geëvalueerd worden in een korte tijdspanne. Als er veel items moeten gecontroleerd worden zou deze fase de vorm van een aparte vergadering kunnen aannemen.

De volgende fase van de vormgeving omvat een meer gedetailleerd overzicht van elk item of van elke taak. Het is belangrijk dat de auteurs niet zelf hun product controleren. Op een school bijvoorbeeld, zouden de leerkrachten die toetsen schrijven voor hun eigen klassen de items van collega's kunnen controleren en op die manier elkaar helpen.

Als er meer dan vier of vijf mensen in een vormgevingsgroep zetelen, wordt de werking van deze groep vaak vertraagd, terwijl minder dan drie personen niet genoeg variëteit brengen in de punten die moeten worden nagezien. Als er meerdere vergaderingen nodig zijn, zou een iemand binnen de organisatie kunnen worden aangeduid als coördinator. Hij zou vergaderingen kunnen plannen, beslissen wie ervoor wordt uitgenodigd en welke materialen er tijdens de vergadering zullen worden besproken.

Deelnemers kunnen op voorhand materialen nakijken om tijd te sparen en tijdens de vergadering kan er dan als volgt worden gewerkt:

- Op tekst gebaseerde items zouden moeten gelezen worden vooraleer de tekst wordt gelezen, want dit helpt items te identificeren die kunnen worden beantwoord zonder de tekst te begrijpen (bv. enkel op basis van gezond verstand of achtergrondkennis).
- Alle andere items kunnen worden beantwoord zonder naar de sleutel te kijken, zoals dit ook tijdens een afname zou gebeuren. Dit zal helpen om te detecteren welke items meer dan een correct antwoord toelaten, welke opties onduidelijk of slecht geformuleerd zijn, welke afleiders twijfelachtig zijn en welke items te moeilijk of onduidelijk zijn.
- Zowel de lengte, stijl, geschiktheid van het onderwerp, als het taalniveau van lees- en luisterteksten zouden moeten worden nagekeken. Het controleren van het taalniveau moet door deskundigen worden gedaan. Hierbij kan worden verwezen naar linguïstische beschrijvingen.

Als de vormgeving in groep gebeurt, kunnen alle problemen die met de materialen vastgesteld worden in detail besproken worden binnen de groep. Vaak is er nogal wat discussie over materialen en itemschrijvers moeten in staat zijn om opbouwende kritiek te aanvaarden en te geven, wat soms moeilijk kan zijn. Als een itemschrijver het nodig vindt om aan ervaren collega's te verantwoorden en uit te leggen waarom hij iets heeft geschreven, dan is het waarschijnlijk dat dit materiaal niet helemaal deugt.

Een persoon in de groep zou de verantwoordelijkheid moeten nemen en een gedetailleerde en adequate neerslag moeten bijhouden van alle beslissingen die over de materialen worden genomen, waarbij duidelijk

aangegeven wordt welke veranderingen bij de vormgeving werden aangebracht. Op het einde van de vergadering is het cruciaal dat het duidelijk is welke veranderingen uiteindelijk werden doorgevoerd.

De toetsaanbieder zou definitieve beslissingen moeten nemen en besluiten wanneer er genoeg discussie is geweest.

De volgende punten zijn van toepassing op vergaderingen die gericht zijn op details:

- Er zou speciale aandacht moeten worden gegeven aan de INSTRUCTIES (die samen met de items aan de kandidaten worden gegeven) en de sleutels.
- Items die op de toets voor bias kunnen zorgen, kunnen geïdentificeerd worden door te refereren aan onderwerpen die moeten worden vermeden of andere aspecten waar voldoende aandacht moet worden aan besteed (zie Bijlage VII).
- Sommige materialen lijken potentieel te hebben, maar hebben meer aanpassingen nodig dan er mogelijk is tijdens de vergadering. Deze worden teruggegeven aan de oorspronkelijke schrijvers om verder te bewerken of ze kunnen aan een meer ervaren schrijver worden toegekend die ze dan verder naziet en vormgeeft.
- Na de vergadering zouden alle reservematerialen en gebruikte toetsen om veiligheidsredenen vernietigd moeten worden. De gewijzigde exemplaren en de aanvaarde materialen worden door de toetsaanbieder bijgehouden.
- Itemschrijvers zouden van de toetsaanbieder feedback moeten krijgen over geweigerde materialen, vooral als ze niet bij de vormgeving betrokken waren, of als ze niet aanwezig waren tijdens de vormgeving van hun eigen materialen.
- Vergaderingen over de vormgeving zijn een uitstekende gelegenheid voor nieuwe itemschrijvers om te leren van het werk van meer ervaren schrijvers door samen te werken in een groep.

### 3.4.2 Pilotstudie, prestest en proefafname

Er is altijd een soort van testfase nodig omdat kandidaten op heel onverwachte manieren kunnen reageren op items. Daarom wordt er een pilot, PRETEST of PROEFAFNAME georganiseerd, afhankelijk van de doelstellingen en de bronnen waarover de toetsaanbieder beschikt.

We spreken van een pilot wanneer er aan een beperkt aantal mensen wordt gevraagd om items te beantwoorden zoals ze dit op een toets zouden doen. Dit kan heel informeel gebeuren en zou, als er niemand anders bereid wordt gevonden, bijvoorbeeld door collega's kunnen worden gedaan. Hun RESPONS wordt daarna geanalyseerd en samen met hun opmerkingen kan deze informatie gebruikt worden om de items verder te verbeteren (zie Bijlage VI).

Pretesten worden meestal georganiseerd voor toetsen die gebruikmaken van items die OBJECTIEF GECORRIGEERD worden. Een pretest wordt afgenomen in dezelfde omstandigheden als een live toets, bij kandidaten die representatief zijn voor de beoogde populatie. Er worden voldoende prestaties verzameld om een statistische analyse mogelijk te maken (zie Bijlage VII). De analyse kan uitwijzen hoe goed de afleideropties hebben gewerkt, hoe moeilijk een item was, wat de GEMIDDELDE score was, of de toets op het correcte niveau was voor de groep kandidaten die de toets hebben afgelegd, hoeveel meetfout er wordt gevonden, of items misschien bias vertonen (zie Bijlage VII), of de items bijdragen tot het meten van hetzelfde construct, hun relatieve moeilijkheidsgraad en andere informatie. De meest elementaire vormen van statistische analyse (zie Bijlag VII) kunnen al ontzettend veel informatie verschaffen en kunnen uitgevoerd worden met goedkope, gebruiksvriendelijke software.

Prestaties op SUBJECTIEF GECORRIGEERDE taken (die schrijven en spreken toetsen) kunnen statistisch worden geanalyseerd, maar een kwalitatieve analyse gebaseerd op een minder groot aantal responsen kan meer informatie opleveren. Kleinschalige pretesting van productieve taken wordt soms *proefafnames houden* genoemd om het te onderscheiden van de pretesting van objectief gecorrigeerde toetsen. Het kan aantonen of de toetstaken goed genoeg hebben gewerkt en de verwachte prestatie hebben uitgelokt.

Anders dan een pilot, vereist een pretest grotendeels dezelfde variabelen als een live toets. Deze omvatten:

- voldoende kandidaten (zie bijlage vii);

- veilig geprinte toetsformulieren;
- toetslokalen en personeel;
- correctoren.

Kandidaten die deelnemen aan een pretest zouden zoveel mogelijk moeten gelijken op de kandidaten die de uiteindelijke live toets zullen afleggen. Een goede oplossing, als dit praktisch haalbaar is, is individuele leerders rekruteren die zich aan het voorbereiden zijn op een live toets.

Feedback beloven op de prestatie is een goede manier om kandidaten te motiveren om deel te nemen en zorgt ervoor dat de responsen echt een afspiegeling zijn van de vaardigheden van de kandidaten. Feedback kan kandidaten en hun leerkrachten helpen om hun huidige vaardigheidsniveau te begrijpen en te bepalen aan welke domeinen ze nog moeten werken vooraleer de live toets kan worden afgelegd.

Een mogelijk nadeel van deze aanpak is dat de items op deze manier openbaar kunnen worden. Dit kan de veiligheid van een toekomstige live toets in het gedrang brengen. Voor sommige toetsaanbieders kan dit ervoor zorgen dat elke vorm van pretesting moeilijk uit te voeren is.

Om het risico te beperken, zouden de toetsformulieren tijdens de pretest niet in dezelfde vorm mogen worden aangeboden zoals dit tijdens de live toets het geval is. Het SAMENSTELLEN VAN EEN TOETS zou goed moeten worden gepland zodat er voldoende tijd wordt voorzien tussen de pretest van een item en het gebruik ervan in een live toets. Als de afname van een prestest extern gebeurt, dan moet het betrokken personeel instructies krijgen over hoe de pretest veilig kan worden gebruikt. Daarnaast moeten betrokkenen een geheimhoudingsverklaring ondertekenen.

Een formulier voor een prestest moet niet erg lijken op de formulieren die tijdens de live toets zullen worden gebruikt, omdat de items worden gepretest, niet de toets zelf. Het kan echter zijn dat de kandidaten die de pretest afleggen, enthousiast zijn als ze weten dat ze de kans krijgen om een toets te maken die heel vergelijkbaar is met de live toets (omdat ze het zien als een manier om zich voor te bereiden). In dit geval is het aan te bevelen om toch een format te gebruiken dat erg lijkt op dat van de live toets.

In alle gevallen zou een pretest onder de omstandigheden van een live toets moeten worden afgenomen. Als de kandidaten verstrooid zijn, spieken of de voorziene tijd wordt niet gerespecteerd, dan kan de data die uit de pretest komen, moeilijk te interpreteren zijn.

Als de kwaliteit van de statistische informatie heel belangrijk is (bv. om de items te KALIBREREN – zie Bijlage VII), dan zouden er, overeenkomstig met de doelstellingen, voldoende kandidaten moeten deelnemen. Het minimum aantal zal afhangen van het soort analyse dat zal worden gebruikt. Toch kan ook een beperkte groep kandidaten (minder dan 50) zinvolle informatie geven die kan aangeven of er problemen zijn met bepaalde items. Met een kleinere groep proefpersonen, zal een kwalitatieve analyse zinvoller zijn.

Het is ook heel belangrijk om kandidaten te vinden die zo representatief mogelijk zijn voor de kandidaten die de live toets zullen maken. Kleinere, minder representatieve groepen kandidaten kunnen leiden tot interpretatiefouten, die dan later gecompenseerd moeten worden tijdens de controle van de items.

Als de pretest wordt uitgevoerd om kwalitatieve informatie over de items te verzamelen, dan moeten een aantal aandachtpunten in acht worden genomen om zoveel mogelijk bruikbare informatie te vergaren:

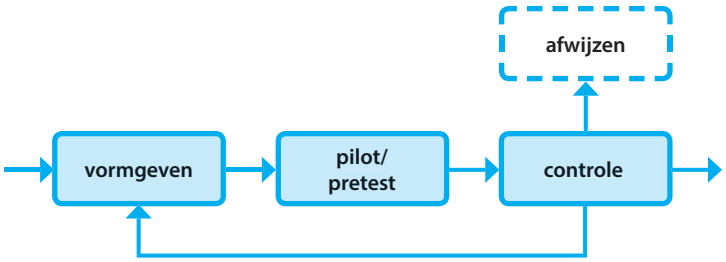
- Voor objectief gecorrigeerde items kan het ook mogelijk zijn om feedback van kandidaten en leerkrachten te verzamelen. Om dit te vereenvoudigen, kunnen er vragen of vragenlijsten worden gebruikt (zie Bijlage VI).
- Voor taken in mondelinge toetsen waarbij er een gesprekspartner nodig is, kan het zinvol zijn om feedback te verzamelen. Zo kan de toetsaanbieder weten of de taak door de kandidaten werd begrepen, geschikt was voor hun ervaring en leeftijdsgroep en of de kandidaten over voldoende informatie beschikten om de taak adequaat uit te voeren (zie Bijlage VI).
- Voor items en taken die subjectief worden gecorrigeerd, kunnen de responsen van de kandidaten aantonen of de kandidaten genoeg mogelijkheden hebben gekregen om te tonen wat het bereik is van de structuur van hun mondeling uiting, de syntactische structuur en de woordenschat die op dit niveau wordt verwacht.
- Er kan ook feedback van kandidaten verzameld worden over hun ervaringen met de pretest of de sessie in zijn geheel (zie Bijlage VI).

### 3.4.3 Evalueren van de items

Na een pilot of pretest zou er een evaluatievergadering moeten volgen met zowel de toetsaanbieder, als ervaren itemsschrijvers en, in het geval van subjectief gecorrigeerde items en taken (bv. schrijftoetsen) ook een ervaren beoordelaar.

Het doel van de vergadering is het bewijs te gebruiken dat tijdens de pilot of pretest werd verzameld om te beslissen welke items worden behouden en welke worden verbeterd of verwijderd.

Dit wordt in Figuur 8 geïllustreerd waar items die moeten worden verbeterd voor de tweede keer aan een pilot of pretest worden onderworpen.




**Figuur 8** ***Items verbeteren door kwaliteitsgarantie***

Wanneer de pretest wordt geëvalueerd, worden volgende algemene punten besproken:

- Welk materiaal is er klaar om in een live toets te worden gebruikt?
- Welk materiaal zou moeten worden afgewezen omdat het niet geschikt is?
- Welk materiaal zou kunnen herwerkt worden vooraleer er besloten wordt of het alsnog in een live toets zal worden gebruikt?

Tijdens de evaluatievergadering zou moeten worden besproken:

- In welke mate de kandidaten vergelijkbaar waren met de doelgroep van de live toets. Hiermee kan worden afgetast in hoeverre het bewijs van de statistische analyses kan worden vertrouwd.
- Hoe activerend en toegankelijk de taken en onderwerpen waren en of er problemen waren met de organisatie.
- Wat de kwaliteit was van de individuele items en taken. Als er subjectief gecorrigeerde taken worden geëvalueerd, dan is het zinvol om over een aantal responsen van kandidaten te beschikken ter controle. Voor de objectief gecorrigeerde items kunnen de statistische analyses problemen aan het licht brengen die dan door deskundige controle kunnen worden bevestigd en gecorrigeerd. Als de analyse echter is gebeurd op basis van twijfelachtige data (bv. ongeschikte kandidaten, of een te beperkte groep), dan moet er heel voorzichtig worden mee omgesprongen. Andere informatie, zoals de inschatting van de kwalitatieve waarde van items en taken, zou dan meer gewicht moeten krijgen.
- Als de statistische en andere informatie vergeleken wordt met taken in een itembank, zou er een coherente en consistente methode moeten worden gevolgd. Dit is ook nuttig tijdens de fase van het samenstellen van een toets. Zie Bijlage VII voor meer informatie over statistische analyses.

## 3.5 Een toets opbouwen

Eenmaal er voldoende materialen beschikbaar zijn, kan de toets worden opgebouwd. De bedoeling van deze fase is om versies van een toets samen te stellen die beantwoorden aan de kwaliteitsstandaarden en eisen van de toetsspecificaties.

Tijdens de fase van het SAMENSTELLEN VAN EEN TOETS worden verschillende aspecten in evenwicht gebracht, zoals de inhoud van de toets en de moeilijkheidsgraad van een item zodat de toets in zijn geheel aan de vereiste toetsspecificaties voldoet.

Sommige kenmerken van de toets worden bepaald door de toetsspecificaties en het format (bv. het aantal en soort items/taken die de toets moet bevatten). Andere kenmerken kunnen flexibel variëren binnen de vastgelegde grenzen (bv. onderwerpen, variatie in accenten, enz.). Richtlijnen kunnen helpen om een geschikte balans te vinden tussen de volgende kenmerken:

- De moeilijkheidsgraad - dit kan subjectieve oordelen bevatten, of wanneer er met een itembank wordt gewerkt, kan de GEMIDDELDE moeilijkheidsgraad van de toetsitems worden beschreven en het BEREIK van moeilijkheid – zie Bijlage VII.
- Inhoud - thema's en onderwerpen.
- Bereik - representativiteit van de taken in relatie tot het construct.
- De progressieve moeilijkheidsgraad - of een toets moeilijker wordt naarmate het aantal taken toeneemt.

Deze richtlijnen zouden voor de gehele toets moeten gelden. Er moet over de onderdelen heen gekeken worden en toetsonderdelen moeten met elkaar worden vergeleken.

Voor sommige soorten toetsen kunnen er aanvullende richtlijnen van toepassing zijn. Bij een leestoets die bijvoorbeeld uit meerdere teksten en items bestaat, zou er moeten worden nagegaan of er geen overlap is tussen de verschillende onderwerpen van de teksten. Ook de lengte van de teksten, uitgedrukt in het totale aantal woorden, moet worden gecontroleerd. In een luistertekst is het dan weer belangrijk om te controleren of er een evenwicht is tussen mannelijke en vrouwelijke stemmen en regionale accenten (als dit relevant is).

## 3.6 Kernvragen

- Hoe zal het ontwikkelingsproces van toetsmaterialen georganiseerd worden?
- Kan er een itembank worden gebruikt?
- Wie zal de materialen schrijven?
- Aan welke professionele eisen zouden de itemschrijvers moeten voldoen?
- Welke training zal er worden gegeven?
- Wie zal er worden betrokken bij de vergaderingen over de vormgeving?
- Is het mogelijk om materialen te PRETESTEN of PROEFAFNAMES te organiseren?
- Wat zouden de consequenties kunnen zijn van een pretest of proefafname en hoe wordt er hier dan mee omgegaan?
- Welk soort analyse zal uitgevoerd worden op de data die tijdens de pretest werden verzameld?
- Hoe zal de analyse worden gebruikt? (bv. voor het SAMENSTELLEN VAN EEN TOETS, voor de opleiding van schrijvers van toetsen)
- Wie zal er betrokken worden bij het samenstellen van de toets?
- Welke variabelen moeten overwogen en tegen elkaar afgewogen worden? (bv. moeilijkheidsgraad, inhoudelijke onderwerpen, bereik van het soort items, enz.)
- Wat zal de rol zijn van de statistische analyse? (bv. bij het bepalen van het gemiddelde en de moeilijkheidsgraad van de toets)
- Als er beslissingen worden genomen, hoe belangrijk zal de statistische analyse dan zijn in relatie tot de andere informatie?
- Zal de toets onafhankelijk worden DOORGELICHT?
- Hoe zal de toets gelinkt worden aan andere versies van dezelfde toets of in een grotere reeks van toetsen worden ingepast?

## 3.7 Aanbevolen literatuur

Voor richtlijnen voor itemschrijvers, zie ALTE (2005).

Voor de analyse van toetstaken, zie ALTE (2004 a, b, c, d, e, f, g, h, i, j, k).

Er zijn taalkundige beschrijvingen beschikbaar van sommige talen die gerelateerd zijn aan het ERK: *Reference Level Descriptors (RLD's)* (Beacco en Porquier, 2007, 2008; Beacco en Porquier, 2004; Glaboniat, Müller, Rusch, Schmitz en Wertenschlag, 2005; Instituto Cervantes, 2007; Spinelli en Parizzi, 2010; **www.englishprofile.org**).

*Threshold* (van Ek en Trim, 1991), *Waystage* (van Ek en Trim, 1990), en *Vantage* (van Ek en Trum, 2001) zijn voorlopers van de RLD's.

Bijlage VII bevat meer informatie over de manier waarop statistische informatie kan worden gebruikt wanneer een toets wordt samengesteld.

# 4 Een toets afnemen

## 4.1 Doel van de afname

Het belangrijkste doel van het afnameproces is accurate informatie verzamelen over de vaardigheid van elke kandidaat.

De grote uitdagingen van het afnameproces zijn van logistieke aard en hebben niet zozeer te maken met de verbetering van de kwaliteit van de toetsmaterialen zoals dit in de vorige stadia het geval was. Toetsaanbieders moeten ervoor zorgen dat:

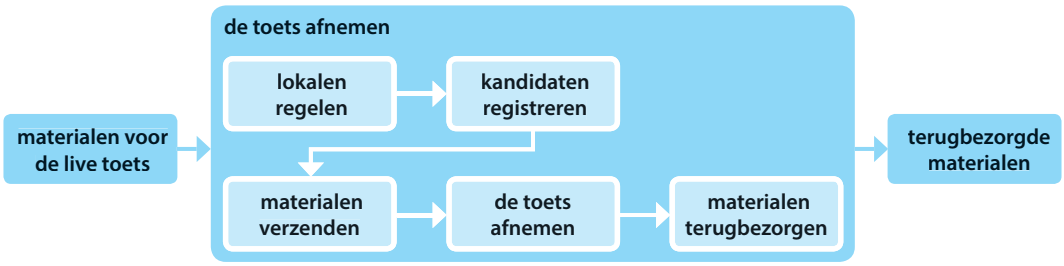
- de prestatie van de kandidaten op de toets zoveel mogelijk beïnvloed wordt door hun taalvaardigheid en zo weinig mogelijk door irrelevante factoren zoals omgevingsgeluid of spieken;
- de RESPONSEN van kandidaten efficiënt en veilig verzameld worden en beschikbaar zijn voor de volgende fases van correctie en beoordeling;
- alle toetsmaterialen op de juiste plaats en op het juiste tijdstip aankomen.

Deze werkzaamheden zijn belangrijk, zowel bij toetsen die op kleine als op grote schaal worden afgenomen. Zoiets eenvoudigs als vooraf de ruimte inspecteren waar de toets zal plaatsvinden, kan een belangrijk verschil maken.

Meer informatie verzamelen over de achtergrond van de kandidaten zou een aanvullende doelstelling kunnen zijn. Vooral wanneer kandidaten niet gekend zijn door de toetsaanbieder kan dit belangrijk zijn. Deze informatie kan inzicht geven in wie er precies deelneemt aan de toets en kan dus bijdragen aan het validiteitsbewijs (zie Bijlage I en Bijlage VII).

## 4.2 Het afnameproces

Het afnameproces wordt voorgesteld in Figuur 9. Verschillende fases, zoals de registratie van de kandidaat of de levering van materiaal, lijken misschien heel eenvoudig in sommige contexten, zoals bij het afnemen van een toets in een klascontext. Toch moet erover gewaakt worden dat ruimtes geschikt zijn voor toetsafnames en dat externe factoren zoals geluid beperkt worden. In andere contexten zal de logistiek een aanzienlijke uitdaging betekenen en deze fases zullen met een grotere inspanning tot een goed einde moeten worden gebracht.




**Figuur 9** ***Het afnameproces***

### 4.2.1 Ruimtes regelen

De ruimtes waarin toetsen worden afgenomen, zouden vooraf moeten worden gecontroleerd. Dat kan door de toetsaanbieder zelf gebeuren of door derden, zoals een vertrouwenspersoon op de school die de toets organiseert. Als er andere organisaties bij betrokken worden, dan zouden die moeten worden goedgekeurd. Ze kunnen geëvalueerd worden aan de hand van deze criteria:

- de capaciteit om het verwachte aantal kandidaten te kunnen verwerken;
- de toegang tot geschikte ruimtes;
- de veiligheid van opslagruimtes;
- de bereidheid om de regels van de toetsaanbieder over te nemen;
- de bereidheid om personeel op te leiden zodat ze de procedures van de toetsaanbieder kunnen volgen.

Als sommige centra door derden worden geleid, zou de toetsaanbieder moeten overwegen om een systeem uit te werken waarbij willekeurige controles kunnen plaatsvinden om de kwaliteit van de afname te controleren die in zijn naam wordt uitgevoerd.

Telkens als er ruimtes gecontroleerd worden, zouden dezelfde criteria moeten worden gehanteerd. Het is waarschijnlijk beter om de ruimtes te controleren vlak voor elke afname, om goed te kunnen inspelen op aspecten die de toetsafname kunnen beïnvloeden.

Dit moet gecontroleerd worden:

- omgevingsgeluid;
- interne akoestiek (vooral voor luistertoetsen);
- grootte (de mogelijkheid om het vereiste aantal kandidaten te laten plaatsnemen, met voldoende ruimte tussen de tafels);
- de vorm van de ruimte (zodat alle TOEZICHTHOUDERS alle kandidaten duidelijk kunnen zien);
- toegankelijkheid;
- de aanwezigheid van faciliteiten zoals toiletten of wachtruimtes voor de kandidaten;
- opslagruimte waar de toetsmaterialen zowel voor als na de toets veilig kunnen worden bewaard.
-

Ruimtes die niet geschikt blijken, of organisaties die bij de afname ernstige fouten maken, kunnen van de lijst van mogelijke ruimtes of partners worden geschrapt.

### 4.2.2 Kandidaten registreren

Gaat het om een toetsafname in een klascontext, dan kan het voldoende zijn om over een lijst met de namen van de studenten te beschikken en hen persoonlijk te kennen. Zijn er voor de toetsaanbieder onbekende kandidaten bij, of kunnen er extra kandidaten aansluiten, dan zou er over deze kandidaten informatie moeten worden verzameld. Een registratieproces verschaft de informatie die nodig is om de toets af te nemen en de resultaten te verwerken en af te leveren. Kandidaten kunnen op dit moment ook om een aangepaste procedure vragen als ze een bepaalde beperking hebben zoals:

- doof of slechthorend zijn;
- slechtziend of blind zijn;
- dyslexie hebben;
- een lichamelijke beperking hebben.

De vraag naar een aangepaste procedure zou op een correcte manier moeten worden behandeld en als dit opportuun is, zou er kunnen worden gezorgd voor een vorm van begeleiding of compensatie. Daarom is het beter dat er standaardprocedures zijn voor de meest voorkomende vragen. Deze zouden het volgende moeten bevatten: het soort bewijs dat de kandidaat moet voorleggen (bv. een doktersbriefje), de acties die ondernomen kunnen worden en de datum waarop de vraag moet zijn ontvangen.

Voor sommige beperkingen, zoals een kandidaat die niet kan lopen en hulp nodig heeft om plaats te kunnen nemen in de ruimte waar de toets plaatsvindt, is het makkelijk om hulp te bieden.

Toch moeten ook andere maatregelen goed overwogen worden. Voor kandidaten die moeilijkheden hebben met lezen, zoals dyslectici of slechtzienden, zouden aangepaste toetsformats of speciale hulpmiddelen

moeten worden voorzien. Zo'n aangepaste afname zou kandidaten niet mogen bevoordelen ten opzichte van anderen.

Het is ook mogelijk om in deze fase achtergrondinformatie te verzamelen over de kandidaat. Kenmerken van de kandidaat kunnen worden gebruikt om belangrijke conclusies te trekken over de mate waarin de kandidaten die de toets afleggen kunnen worden vergeleken. Dit kan gaan om:

- scholingsgraad;
- moedertaal;
- geslacht;
- leeftijd;
- ervaring met het leren van de doeltaal.

In alle gevallen moet de reden waarom deze informatie gevraagd wordt duidelijk zijn voor de kandidaten. Alle persoonlijke materiaal dat verzameld wordt moet bovendien veilig bewaard worden zodat de privacy van de kandidaten beschermd wordt.

Het registratiemoment is niet alleen goed om informatie te verzamelen, maar kan ook aangegrepen worden om de kandidaten te informeren. Zij zouden moeten vernemen welke voorwaarden aan de registratie verbonden zijn, welke regels er gelden tijdens de afname, hoe eventueel bezwaar moet worden ingediend, hoe en wanneer ze een speciale procedure kunnen aanvragen,… Kandidaten moeten weten in welke ruimte en hoe laat ze de toets zullen afleggen en misschien kan er hen op dat moment nog praktische informatie worden gegeven. Om ervoor te zorgen dat alle kandidaten volledig en correct worden geïnformeerd, kan de informatie in geprinte vorm, op de website of via een standaard e-mail doorgegeven worden.

De registratie kan gedaan worden door de toetsaanbieder zelf, door wie de toets afneemt of door derden. De toetsaanbieder moet er wel voor zorgen dat de registratie voor alle kandidaten zo uniform mogelijk verloopt.

### 4.2.3 Materiaal verzenden

Misschien zullen de materialen moeten worden geleverd in de ruimtes waar de toets wordt afgenomen. Dat moet dan veilig en op tijd gebeuren zodat alles klaar is voor de afname begint.

Het is vaak beter om de materialen een heel eind voor de afname plaatsvindt te versturen zodat ze zeker op tijd ter plaatse zijn en eventuele verloren gegane stukken kunnen worden vervangen. Het is dan wel belangrijk dat de toetsaanbieder er kan op rekenen dat de materialen, zolang ze zich op de plaats van de afname bevinden, veilig worden bewaard.

Wie de toets afneemt, zou aan de hand van een lijst moeten controleren of hij ook werkelijk de vereiste materialen heeft ontvangen. Als er iets ontbreekt of er als er schade wordt vastgesteld, dan zou er een procedure moeten worden gevolgd die bepaalt hoe vervangende of aanvullende materialen kunnen worden gevraagd.

### 4.2.4 De afname van een toets

Voor de dag van de afname moeten er voldoende toezichthouders, beoordelaars en andere ondersteunende personeelsleden voorzien worden. Iedereen die bij de afname betrokken is, zou op voorhand zijn verantwoordelijkheden moeten kennen. Als er veel personeelsleden en ruimtes nodig zijn, zou iedereen via een rooster kunnen worden geïnformeerd.

De richtlijnen voor de afname van de toets zouden ook instructies moeten bevatten die beschrijven hoe de identiteitsdocumenten van kandidaten kunnen worden gecontroleerd en wat er met laatkomers moet gebeuren.

Voor de toets begint, zouden de kandidaten duidelijke instructies moeten krijgen over hoe ze zich tijdens de afname moeten gedragen. Dit kan informatie zijn over ongeoorloofde materialen, het gebruik van mobiele telefoons, de ruimte verlaten tijdens de afname en het begin- en einduur van de afname. Ze zouden ook moeten worden gewaarschuwd voor onaanvaardbaar gedrag zoals influisteren en overschrijven.

De toezichthouder die tijdens de afname aanwezig is, zou goed moeten weten wat hij moet doen als de regels worden overtreden, of als er dingen voorvallen die al dan niet konden worden voorzien, bv. kandidaten die spieken of stroomuitval. Er kan ook iets anders gebeuren waardoor er bias of ongelijkheid ontstaat en dan moet de toezichthouder de afname misschien stopzetten. Wanneer er wordt gespiekt, zou de toezichthouder zich bewust moeten zijn van de mogelijke gevaren die digitale toestellen zoals opnameapparatuur, MP3-spelers, pennen waarmee gescand kan worden en mobiele telefoons met camera's met zich meebrengen.

Als het om onvoorziene omstandigheden gaat, kan de toezichthouder worden gevraagd om zelf een oordeel te vellen, dat hij daarna aan de toetsaanbieder rapporteert. Dit rapport zou in detail moeten vermelden hoeveel kandidaten werden benadeeld, wanneer het probleem werd vastgesteld en wat er gebeurd is. Er kan aan de toezichthouder ook een telefoonnummer worden bezorgd zodat hij in geval van nood om advies kan vragen.

### 4.2.5 Terugbezorgen van materiaal

De toetsmaterialen zouden gebundeld en terugbezorgd moeten worden aan de toetsinstelling of vernietigd moeten worden. Als ze teruggestuurd worden, kan er direct na de afname ook andere relevante documentatie worden meegezonden zoals aanwezigheidsregisters en een plattegrond van de ruimtes. Materialen zouden op een veilige manier moeten worden terugbezorgd. Hiervoor zou gebruik kunnen worden gemaakt van dezelfde dienst die de materialen ook had geleverd voor de afname. Deze dienst zou de materialen op elk moment moeten kunnen opsporen, voor het geval er verzendingen vertraging oplopen of verloren gaan.

## 4.3 Kernvragen

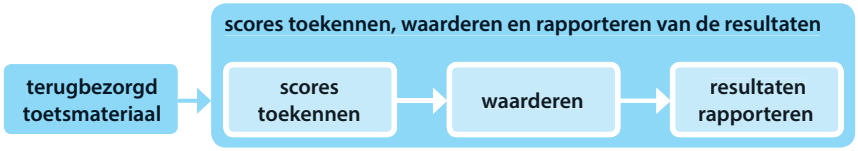
- Welke middelen zijn er beschikbaar voor de afname van de toets? (administratief personeel, toezichthouders, ruimtes, cd-spelers, enz.)
- Hoe moet het personeel worden getraind?
- Hoe kunnen middelen zoals ruimtes en cd-spelers gecontroleerd worden voor de dag waarop de toets wordt afgenomen?
- Met welke frequentie worden er toetssessies aangeboden?
- Hoeveel kandidaten worden er verwacht?
- Hoe zullen kandidaten geregistreerd worden en hoe zal hun aanwezigheid op de toets vastgelegd worden?
- Hoeveel ruimtes zullen er gebruikt worden? Als er meerdere ruimtes gebruikt worden, liggen die dan ver van elkaar of zijn ze moeilijk te bereiken?
- Hoe zullen materialen van en naar de afnamecentra verzonden worden?
- Hoe zullen de toetsmaterialen veilig bewaard worden voor de afname?
- Wat zou er mis kunnen gaan? Zijn er procedures en regels die bepalen wat er dan zou moeten gebeuren?

## 4.4 Aanbevolen literatuur

Zie ALTE (2006b) voor een checklist voor zelfevaluatie die gebruikt kan worden voor de logistiek en afname.

# 5 Scores toekennen, waarderen en resultaten rapporteren

Het doel van scores toekennen is om de prestatie van elke kandidaat op een accurate en betrouwbare te evalueren. WAARDEREN houdt in dat elke kandidaat in een betekenisvolle categorie wordt ingedeeld zodat zijn toetsscore makkelijker begrepen kan worden. Een betekenisvolle categorie zou een van de niveaus van het ERK kunnen zijn, zoals A2 of C1. Wanneer de resultaten worden gerapporteerd, is het de bedoeling dat kandidaten en andere BELANGHEBBENDEN de toetsresultaten ontvangen en alle informatie die nodig is om deze op een gepaste manier te gebruiken; zoals een beslissing nemen over een kandidaat, bijvoorbeeld hem wel of geen job aanbieden. Figuur 10 geeft een overzicht van een gangbare versie van dit proces. In sommige gevallen kan de evaluatie van de prestatie van kandidaten echter gebeuren tijdens de toets zelf. Spreekvaardigheid bijvoorbeeld wordt soms op deze manier geëvalueerd, hoewel de scores aangepast kunnen worden door de toetsaanbieder voor de resultaten gerapporteerd worden.




**Figuur 10** ***Het proces van scores toekennen, waarderen en rapporteren van de resultaten***

### Voorafgaande stappen

Voor er met scores toekennen en waarderen wordt begonnen, moeten volgende stappen worden ondernomen:

- een aanpak voor het toekennen van scores ontwikkelen;
- CORRECTOREN en BEOORDELAARS rekruteren;
- correctoren en beoordelaars trainen.

## 5.1 Scores toekennen

Hoewel de uitdrukking 'scores toekennen' alle activiteiten omvat die te maken hebben met het toekennen van een score aan RESPONSEN op toetsonderdelen, wordt er meestal een onderscheid gemaakt tussen de *corrector*, van wie minder deskundigheid wordt vereist, en de *beoordelaar*, die professioneel opgeleid moet zijn. Dit is het onderscheid dat we ook maken in deze tekst. Dit onderdeel gaat over automatische correctie (door een persoon) en computercorrectie.

### 5.1.1 Methodische correctie

METHODISCHE CORRECTIE moet niet gebeuren door toetsdeskundigen. Wie de taal die wordt getoetst op een hoog niveau beheerst, komt al in aanmerking als corrector. Toch hebben ook de correctoren training nodig en moeten ze begeleid worden en over eenduidig interpreteerbare correctiesleutels beschikken om goed werk te kunnen leveren. Als de correctie door een kleine groep collega's gebeurt, kunnen zij de kwaliteit van elkaars werk beoordelen.

Het correctieproces moet zo uitgevoerd worden dat de procedures zoals gepland worden gevolgd en de resultaten beschikbaar zijn op het afgesproken tijdstip. De werkdruk mag echter niet zo hoog zijn dat de betrouwbaarheid en de accuraatheid van de correctie in het gedrang komen.

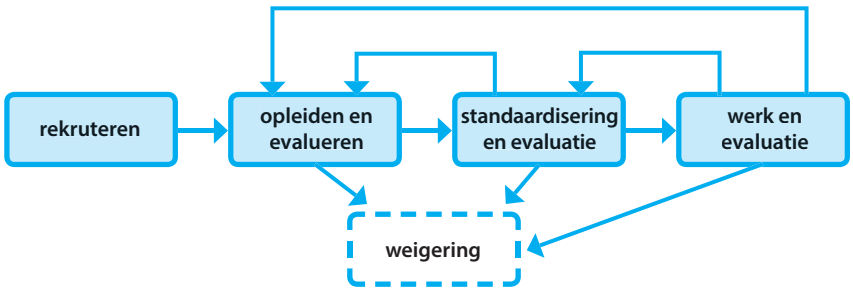
### De rekrutering en training van correctoren

De meest eenvoudige manier van methodische correctie koppelt de respons van een kandidaat aan één of meerdere vastgelegde antwoorden. Meerkeuze-items zijn hiervan het duidelijkste voorbeeld. Zij laten geen ander antwoord toe dan de opties die worden gegeven. Als dit soort correctie vereist is, dan moet de corrector de taal beheersen die getoetst wordt. Hij moet oog hebben voor details en bereid zijn om repetitieve handelingen te verrichten. Andere vaardigheden worden niet verwacht. Tijdens hun opleiding zou dit soort correctoren zich de procedures eigen moeten maken die moeten worden gevolgd. Gezien de technologie waarover we tegenwoordig beschikken, kan dit soort correctie even goed, of zelfs beter, door een computer gebeuren.

Als bij items de respons niet zomaar aan een vooraf bepaald antwoord kan gekoppeld worden, dan moet de corrector misschien enige kennis hebben van de taal, de taalleerders en het construct van de toets. Want aan PARTIAL CREDIT ITEMS bijvoorbeeld kunnen meerdere scores worden toegekend, afhankelijk van hoe goed de respons was. Er zou bijvoorbeeld een score kunnen worden toegekend wanneer een juist werkwoord wordt geselecteerd en een hogere score wanneer ook de juiste vorm wordt gekozen. Correctoren moeten dan een zeker niveau van relevante deskundigheid hebben om het onderscheid te kunnen maken tussen een correcte en foute respons.

Voor dit soort items kan het ook moeilijk zijn om te garanderen dat de correctiesleutel exhaustief is. Het helpt dus als een corrector alternatieve responsen kan herkennen en rapporteren.

Omdat correctoren tijdelijk worden aangeworven, maar opnieuw in dienst kunnen worden genomen voor de correctie van andere afnames, is het belangrijk om een evaluatiesysteem op te zetten dat gebaseerd is op een aantal parameters zoals accuraatheid, betrouwbaarheid en snelheid. Correctoren die niet voldoen, zouden dan kunnen worden geweigerd of kunnen opnieuw worden getraind. Zo'n systeem kan verbonden worden aan een opleiding, zoals in Figuur 11 wordt voorgesteld. Correctoren die vaak opdrachten krijgen, moeten niet aan alle trainingen deelnemen. Als hun prestatie wordt geëvalueerd, zal het makkelijker zijn om te beslissen of een corrector opnieuw ingezet wordt, eerst bijkomende training nodig heeft of vervangen moet worden.



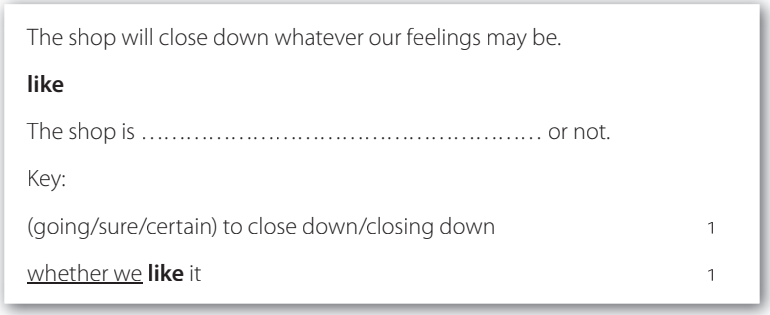

**Figuur 11** ***Rekrutering, opleiding en evaluatie van correctoren***

### Richtlijnen om responsen te evalueren

Een geformaliseerde sleutel is de beste manier om de juiste antwoorden vast te leggen en aan de correctoren te communiceren. Sleutels voor items worden samen met de betreffende items ontwikkeld en volgens dezelfde procedures. De aanvaardbare responsen die de sleutel bevat, zouden zo helder mogelijk moeten zijn uitgeschreven. De correctoren zouden ze ook slechts op één manier mogen kunnen interpreteren.

Figuur 12 toont een voorbeeld van hoe kandidaten gevraagd worden om een leemte in te vullen met het gegeven woord ('like'). De sleutel voorziet vier mogelijke alternatieven voor het eerste element (score 1) en een mogelijkheid voor het tweede (score 1). De maximale score die dus voor dit item kan worden behaald, is 2.

Een duidelijke vormgeving van de sleutel en andere documenten zal ervoor zorgen dat correctoren efficiënter, accurater en betrouwbaarder werk leveren.


The shop will close down whatever our feelings may be.

**like**

The shop is …… or not.

Key:

(going/sure/certain) to close down/closing down 1

<u>whether we</u> **like** it 1

**Figuur 12** ***Een voorbeeld van een item waarbij een stukje tekst moet worden ingevuld***

Er zijn misschien ook andere correcte antwoorden mogelijk, maar die werden niet in de sleutel opgenomen. Daarom zou er aan de correctoren moeten worden gevraagd om alle andere antwoorden die naar hun gevoel juist zijn, vast te leggen. Deze zouden geëvalueerd moeten worden en als ze juist blijken te zijn, dan moeten de kandidaten ervoor beloond worden. Als de groep correctoren klein is, dan kunnen regelmatige discussies met diegene(n) die de toets heeft/hebben opgebouwd afdoende zijn om problemen op te lossen. In andere situaties, als de sleutel geëvalueerd en aangepast wordt, zullen de toetsen opnieuw moeten worden gecorrigeerd.

**Het correctieproces in goede banen leiden**

Normaal is de tijd die voor de correctie is voorzien beperkt. De resultaten moeten immers tegen een bepaalde datum aan de kandidaten worden meegedeeld. De tijd die nodig is, kan worden ingeschat door het aantal kandidaten en beschikbare correctoren in aanmerking te nemen. Het is beter om de tijd lichtjes te overschatten of meer correctoren in te zetten dan eerst voorzien, om het hoofd te kunnen bieden aan onverwachte moeilijkheden.

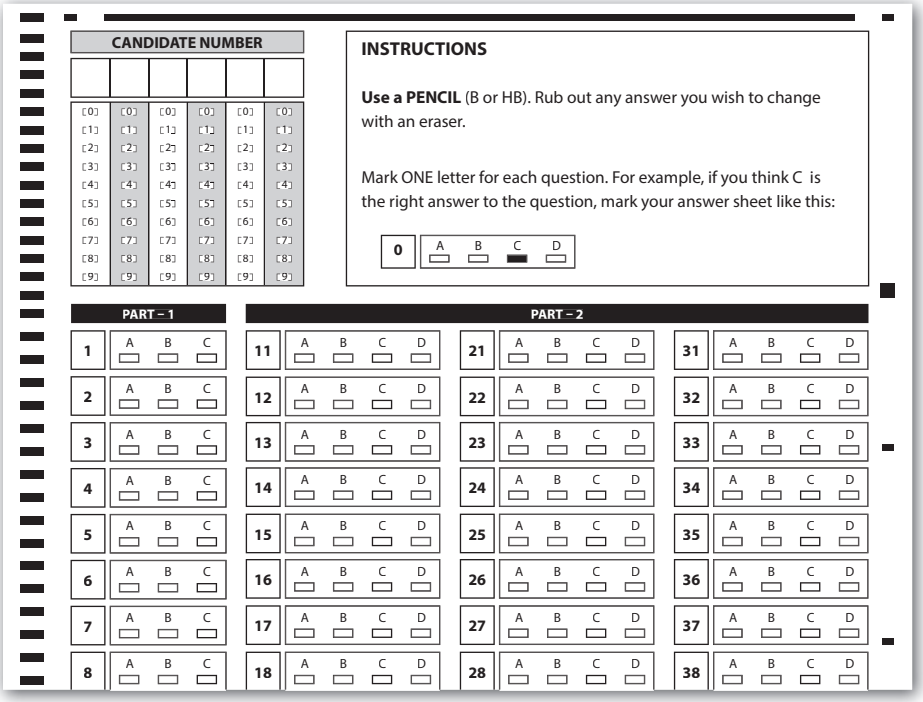
Als het over veel kandidaten en correctoren gaat, is er een systeem nodig om tijdens het proces toetsen te kunnen traceren. Een eenvoudig voorbeeld is een systeem waarbij het nummer van het SCHRIJFPRODUCT bijgehouden wordt met het nummer van de corrector, de datum waarop het werd ontvangen en gecorrigeerd. Zo'n systeem zou de toetsaanbieder moeten helpen in te schatten hoeveel tijd een corrector nodig heeft voor het verbeteren van een bepaald aantal toetsen.

Het traceersysteem kan ook elementaire informatie verschaffen die kan gebruikt worden om elke corrector te evalueren zoals de gemiddelde tijd die hij nodig heeft om een schriftelijke respons te corrigeren. Als het werk van de corrector is geëvalueerd, kan ook worden berekend hoeveel fouten hij gemiddeld maakt. Het zou voldoende kunnen zijn om enkel een representatieve steekproef te nemen van het werk van elke corrector om deze berekeningen uit te voeren.

## 5.1.2 Computercorrectie

Als schriftelijke toetsen met een computer worden gecorrigeerd, dan wordt er meestal gebruik gemaakt van een OPTISCHE LEZER (*optical mark recognition* of OMR). OMR is ontzettend zinvol als er grote aantallen schriftelijke toetsen moeten worden verbeterd en wanneer de items van die aard zijn dat ze geen menselijke beoordeling vereisen (bv. meerkeuze, waar/onwaar of multiple matching). Kandidaten kunnen hun antwoorden registreren op aangepaste OMR-bladen zoals geïllustreerd in Figuur 13. Een OMR-scanner legt dan hun respons vast zodat die daarna op de computer kan worden opgeslagen. OMR-technologie kan ook gebruikt worden voor items die methodische correctie vereisen. De corrector registreert dan de correctie op een OMR-blad dat daarna wordt gescand.

Scanners versnellen het toekennen van scores en verminderen het aantal menselijke fouten, maar het kan gebeuren dat de scanner niet alles leest of ongewilde scores leest. Om zo'n fouten te voorkomen, kunnen integriteitscontroles gebeuren op de data. Dit houdt in dat de OMR-bladen nagekeken worden om responsen te vinden die lijken in te gaan tegen de eisen van de toets zoals meerdere responsen voor een item wanneer er slechts één moet worden geselecteerd. Alle aanpassingen van de OMR-bladen zullen dan handmatig moeten gebeuren.




**Figuur 13** ***Stuk uit een OMR-blad***

### 5.1.3 Beoordelen

Hier zullen we *beoordelen* en *beoordelaar* gebruiken om te refereren aan het toekennen van scores waarvoor, in een grotere mate dan bij de methodische correctie, getrainde deskundigen nodig zijn. Als er een oordeel moet worden geveld, kan de toetsaanbieder voorafgaand aan de beoordeling geen uniek 'goed antwoord' verschaffen. Daarom is er vaak discussie bij beoordelingen, meer dan dit bij andere methodes voor het toekennen van scores het geval is. Bij beoordelingen bestaat het gevaar van inconsistentie tussen verschillende beoordelaars of zelfs binnen het werk van één enkele beoordelaar. Een combinatie van training, monitoring en correctieve feedback kan gebruikt worden om een accurate en betrouwbare beoordeling te verzekeren.

Veel van wat er gezegd werd over de methodische correctie is ook waar voor de beoordeling: het proces zou zo moeten worden uitgevoerd dat de middelen efficiënt worden ingezet. Controles en monitoring zouden ervoor moeten zorgen dat de accuraatheid niet in het gedrang komt. De betrouwbaarheid van beoordelingen zou ook moeten worden gemonitord (zie hoofdstuk 13, Bijlage VII).

#### Beoordelingsschalen

De meeste methodes voor het beoordelen van vaardigheden hangen af van een bepaalde BEOORDELINGSSCHAAL. Dit is een set van descriptoren die prestaties op verschillende niveaus beschrijven. Ze tonen welke score of WAARDERING aan elke prestatie zou moeten toegekend worden.

Beoordelingsschalen beperken de variatie die inherent is aan de subjectiviteit van menselijke oordelen. Er kunnen heel wat opties worden overwogen:

- **Holistische of analytische schalen:** er kan één enkele score worden toegekend aan een prestatie door gebruik te maken van een schaal die elk niveau van prestatie beschrijft, misschien in termen van een reeks van kenmerken. De beoordelaar kiest het niveau dat de prestatie het best beschrijft. Er kunnen ook schalen worden ontworpen voor diverse criteria (bv. communicatieve effect, accuraatheid, beoogde inhoudelijke weergave, enz.) en dan wordt er aan elk van de criteria een score toegekend. Beide methodes kunnen gerelateerd zijn aan eenzelfde construct van taalvaardigheid, dat in dezelfde bewoordingen wordt gedefinieerd: het verschil zit in het oordeel dat de beoordelaar wordt verondersteld te vellen.

- **Relatieve of absolute schalen:** schalen kunnen in relatieve, evaluatieve termen worden beschreven (bv. 'zwak', 'adequaat', 'goed') of ze kunnen tot doel hebben taalvaardigheidsniveaus in positieve, welomlijnde bewoordingen te definiëren. Als we een prestatie in termen van de ERK-schalen en -niveaus wensen te interpreteren, dan geniet deze tweede optie de voorkeur. De ERK-descriptorenschalen kunnen dan dienen om dit soort beoordelingsschalen uit te werken.
- **Schalen of checklists:** een alternatieve of aanvullende methode om met beoordelingsschalen te werken, is scores toekennen op basis van een lijst met 'ja'/'nee'-oordelen waarmee wordt aangegeven of een prestatie voldoet aan specifieke eisen of niet.
- **Generische of taakspecifieke schalen:** een toets kan een generische schaal of een set van schalen voor alle taken hanteren, of beoordelingscriteria voorzien die specifiek zijn voor elke taak. Een combinatie van beide is mogelijk: er kunnen specifieke criteria worden voorzien om de uitvoering van de taak te beoordelen (een lijst met inhoudelijke punten die aan bod moeten komen), terwijl de andere schalen generisch kunnen zijn.
- **Comparatieve of absolute beoordelingen:** het is mogelijk om een schaal te definiëren aan de hand van voorbeeldprestaties. De taak van de beoordelaar is dan niet om een absoluut niveau van prestatie toe te kennen, maar om eenvoudigweg te zeggen of de prestatie slechter of beter of hetzelfde is als in het voorbeeld. Een score is dan een rangorde op een schaal. De interpretatie van deze rangschikking, bv. in termen van de ERK-niveaus, hangt dan af van oordelen over het niveau dat geïllustreerd wordt door de voorbeelden. Deze aanpak zal waarschijnlijk het best werken als de voorbeelden taakspecifiek zijn.

Hoewel deze methodes erg van elkaar lijken te verschillen, hangen ze allemaal af van vergelijkbare onderliggende principes:

- alle beoordelingen hangen af van de manier waarop de beoordelaars de niveaus begrijpen;
- voorbeelden zijn essentieel om deze betekenis te definiëren en erover te communiceren;
- wanneer er met schalen wordt gewerkt, spelen de toetstaken die gebruikt worden om een respons uit te lokken een cruciale rol.

In het verleden hadden de niveaus een betekenis die verbonden was met de context van een specifieke toets en de kandidaten die hem aflegden. Daarom was het moeilijk om er de niveaus van toetsen uit andere contexten mee te vergelijken. De ontwikkeling van referentiekaders voor vaardigheden zoals het ERK biedt de mogelijkheid om niveaus die in lokale contexten worden gebruikt te begrijpen in relatie tot andere contexten. Dit heeft een impact op de manier waarop beoordelingsschalen worden verwoord.

In het verleden was het niveau impliciet en werden schalen vooral uitgedrukt in relatieve, evaluatieve termen. Vandaag worden schalen vooral geformuleerd in termen die voortvloeien uit de herkenbare manier waarop het ERK taalvaardigheidsniveaus beschrijft en waarbij gebruik wordt gemaakt van positieve en welomlijnde stellingen. Dit verandert niks aan het feit dat de voorbeelden (eerder dan de geschreven tekst van de descriptoren) essentieel blijven om het niveau te definiëren en erover te communiceren. Het is echter goed dat toetsaanbieders aangespoord worden om explicieter te omschrijven wanneer een niveau wordt bereikt.

Het ERK promoot het denken en werken in termen van criterium-gerelateerde taalvaardigheidsniveaus. Niveaus worden altijd aan de hand van twee aspecten beschreven: *wat* mensen kunnen en *hoe* goed ze dit kunnen. In een toets wordt het 'wat' gedefinieerd door de taken die worden gespecifieerd. Hoe goed deze taken worden uitgevoerd, is wat de beoordelaars moeten beoordelen.

Dit is de reden waarom de traditionele evaluatieve benadering voor het formuleren van beoordelingsschalen degelijk lijkt te werken, op voorwaarde dat de taken goed gekozen zijn en de oordelen de prestaties op de taken betreffen. Taken zijn dus cruciaal bij het vastleggen van schalen, zelfs als er meer of minder expliciet naar verwezen wordt om te definiëren wat een 'geslaagde' prestatie inhoudt.

**Het beoordelingsproces**

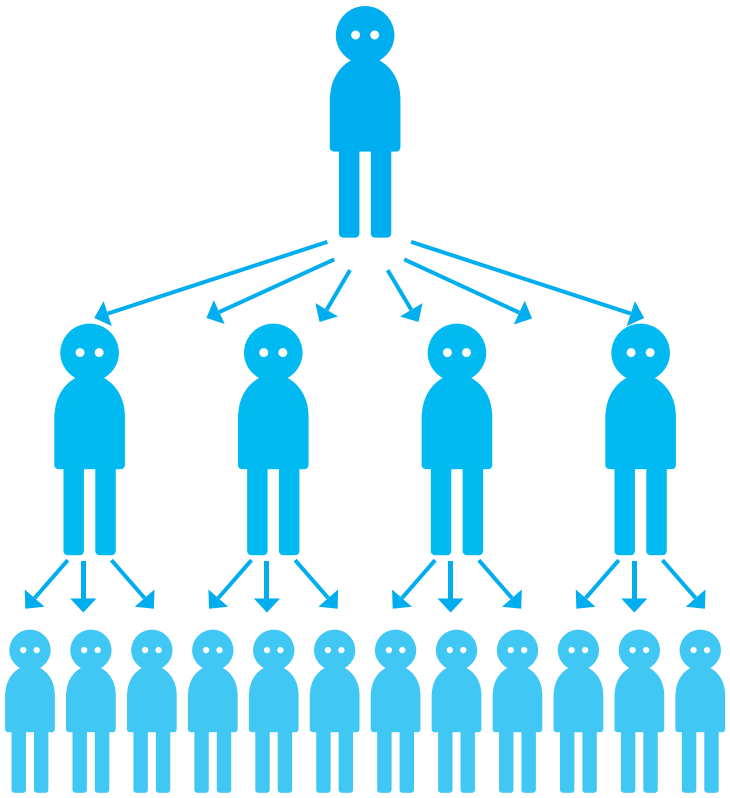
Om het beoordelingsproces vlot te laten verlopen, moeten de beoordelaars een goed begrip hebben van de standaard. De voorbeeldprestaties die worden gedeeld, dienen als basis van dit gedeelde begrip.

Worden toetsen op kleine schaal afgenomen, dan kunnen de beoordelaars tot dit gedeelde begrip komen dankzij een vrije en gelijkwaardige discussie. Dit kan betekenen dat de kandidaten gelijk behandeld worden, maar garandeert niet dat de standaard die wordt overeengekomen meer dan een lokale betekenis heeft of

stabiel blijft over sessies heen. Worden toetsen op grote schaal gebruikt, dan moet de standaard stabiel en betekenisvol zijn. Dit zal in de praktijk afhangen van ervaren toetsontwikkelaars die de autoriteit hebben om de standaard te communiceren aan de nieuwkomers.

Een kleine groep van ervaren beoordelaars zorgt er dus voor dat de belangrijkste pijlers van de standaard overeind blijven: de training, monitoring en de verbetering van andere beoordelaars.

Dit hiërarchische systeem kan uit verschillende niveaus bestaan zoals in Figuur 14 te zien is. Het kan een efficiënte manier zijn om face-to-facetraining te organiseren en het werk van correctoren te monitoren. Door de moderne informatietechnologie en de ontwikkeling van onlinetrainingen is er echter minder nood aan hiërarchie. Het is ook belangrijk om op te merken dat dezelfde gezaghebbende voorbeeldbeoordelingen er op elk niveau voor zorgen dat de overdracht van de standaard accuraat verloopt.



**Figuur 14** ***Een stabiele standaard garanderen door middel van een systeem met teamleiders***

### Beoordelaarstraining

Door de beoordelaars te trainen, streven we een consistente en accurate beoordeling na. Standaardisatie is een proces van beoordelaars trainen om de beoogde standaard toe te passen. Als het ERK als maatstaf wordt gebruikt, dan zou de training moeten beginnen met oefeningen om de beoordelaars kennis te laten maken met het ERK. Er zou gerefereerd kunnen worden aan ERK-gerelateerde illustratieve voorbeelden van prestaties voor spreken of schrijven (Raad van Europa 2009). Het kan ook nodig zijn om beoordelaars te trainen los van de beoordelingsschalen waarmee ze vertrouwd zijn. De training kan dan bestaan uit een reeks stappen, gaande van een open discussie tot een onafhankelijke beoordeling, waarbij de gebruikte voorbeelden gerelateerd zijn aan de toets die wordt beoordeeld:

- geleide discussie over een voorbeeld, waardoor beoordelaars het niveau begrijpen;
- onafhankelijke beoordeling van een voorbeeld gevolgd door de vergelijking met een vooraf toegekende score en diepgaande discussie over de redenen waarom er verschillen zijn;
- onafhankelijke beoordeling van verschillende voorbeelden om te tonen hoe dicht beoordelaars bij de vooraf toegekende score komen.

Waar het mogelijk is, zouden de voorbeelden prestaties moeten zijn die uit de huidige afnamesessie komen. Als dit niet mogelijk is, dan moeten er taken uit vorige sessies worden gebruikt.

**Monitoring en kwaliteitscontrole**

Idealiter zijn de beoordelaars na de training in staat om voldoende accuraat en consistent te werken zodat verdere feedback en bijsturing niet nodig zijn en kan de beoordelingsfase vlot verlopen. Toch blijft monitoring nodig om problemen vroeg genoeg te identificeren en te remediëren.

We kunnen vier soorten problemen of 'beoordelaarseffecten' onderscheiden:

1. **Strengheid of mildheid:** de beoordelaar kent steevast te lage of te hoge scores toe.
2. **Gebruikte correctiebereik (centrale tendentie):** de beoordelaar gebruikt een te beperkt bereik van scores waardoor het onderscheid tussen goede en slechte prestaties niet duidelijk genoeg is.
3. **Halo-effect:** als een beoordelaar meerdere scores moet toekennen, dan laat hij zich beïnvloeden door de eerste score om de andere scores toe te kennen. Die worden dan bepaald onafhankelijk van het werkelijke niveau van de prestatie.
4. **Inconsistentie:** de beoordelaar past de standaard niet systematisch toe zodat de scores niet overeenstemmen met die van de andere beoordelaars.

Hoe ernstig deze problemen zijn, hangt voor een stuk af van hoe er kan worden bijgestuurd. Zo blijken beoordelaars een soort van ingebouwde strengheid te hebben. Als er geprobeerd wordt om hier tegenin te gaan in een poging de standaard te bereiken, dan kan dit ongewenste gevolgen hebben: beoordelaars kunnen onzeker en minder consistent worden. Het kan dus beter zijn om een systematisch niveau van strengheid of mildheid te aanvaarden, toch zolang er een statistische procedure beschikbaar is om dit te corrigeren. Schalering of een itemresponsmodel gebruiken zijn hier twee opties (zie Bijlage VII).

Wanneer er een te beperkt bereik van scores wordt gehanteerd, is dit slechts gedeeltelijk te corrigeren met statistische procedures. Dit probleem moet dus snel geïdentificeerd en geremedieerd worden; ofwel door de beoordelaar te trainen, ofwel door hem te ontslaan.

Een vorm van monitoring is dus wenselijk, al is ditt makkelijker te realiseren in real time in het geval van schrijven, waar een schrijfproduct van beoordelaar aan beoordelaar kan worden doorgegeven. Mondelinge evaluatie is veel moeilijker te monitoren tenzij de mondelinge productie wordt opgenomen. In dit geval zou er voor de start van de beoordelingsfase meer aandacht moeten worden besteed aan de training van de beoordelaars. Het genereren van elementaire informatie over de prestatie van de beoordelaar kan helpen bij dit proces (zie Bijlage VII).

Monitoring gebeurt op veel manieren, gaande van eenvoudige (bv. informele controles en mondelinge feedback aan beoordelaars) tot complexe (bv. gedeeltelijke herbeoordeling van het werk van de beoordelaar en samenvattende gegevens produceren om een indicatie te hebben van de prestatie) procedures. Een aantrekkelijke methode is om vooraf beoordeelde schrijfproducten toe te wijzen aan een beoordelaar en te kijken in welke mate zijn scores overeenstemmen. Om dit betrouwbaar te kunnen doen, moeten deze schrijfproducten onmogelijk te onderscheiden zijn van andere schrijfproducten, dus fotokopies maken is bijvoorbeeld niet mogelijk. In de praktijk is deze methode enkel haalbaar voor schrijfproducten die uit een computergestuurde toets komen, of wanneer schrijfproducten gescand worden voor het online toekennen van scores.

Een andere manier om het aantal beoordelingsfouten terug te dringen, is beoordelaars met elkaar vergelijken (zodat bepaalde beoordelaarseffecten zichtbaar worden en statistisch kunnen worden gecorrigeerd) door DUBBELE BEOORDELING toe te passen of een gedeelte door meerdere beoordelaars te laten beoordelen. Afhankelijk van de statistische procedures die er worden gebruikt, zal er een bepaalde methode nodig zijn om de informatie te combineren en tot een finale score voor de kandidaat te komen.

## 5.2 Waarderen

Het hele proces van toetsen ontwerpen, ontwikkelen, afnemen en scores toekennen zoals het tot dusver werd beschreven, leidt tot het moment waarop we de prestatie van elke kandidaat kunnen evalueren en dit op een bepaalde manier kunnen rapporteren.

In sommige contexten rangschikt een toets enkel kandidaten van hoog naar laag en worden er misschien arbitraire slaaggrenzen opgetrokken tussen beide groepen (bv. de top 10% krijgt Graad A, de volgende 30% krijgt Graad B enz). Hoewel deze norm-gerelateerde aanpak in de maatschappij een belangrijke rol kan spelen, is ze niet bevredigend in die zin dat de prestatie enkel geëvalueerd wordt in relatie tot de andere kandidaten; het zegt niks over wat de prestatie *betekent*, bv. in termen van taalvaardigheid.

De alternatieve, meer betekenisvolle aanpak is de criterium-gerelateerde waarbij de prestatie wordt geëvalueerd ten opzichte van enkele vaststaande, absolute criteria of standaarden. Dit is duidelijk het geval bij taaltoetsen die de resultaten in termen van ERK-niveaus rapporteren.

Een toets kan ontworpen zijn om over verschillende of slechts één ERK-niveau te rapporteren. In het laatste geval zou er tegen de kandidaten die het niveau bereikt hebben, kunnen worden gezegd dat ze geslaagd zijn. De anderen krijgen dan te horen dat ze gezakt zijn. Er kan ook aangegeven worden in welke mate iemand geslaagd of gezakt is.

De score identificeren die overeenkomt met het bereiken van een bepaald niveau, wordt CESUURBEPALING genoemd. Er zijn onlosmakelijk subjectieve oordelen mee verbonden, maar de cesuurbepaling wordt zo goed mogelijk gebaseerd op gegevens.

Verschillende methodes voor cesuurbepaling worden toegepast op productieve vaardigheden (spreken, schrijven) en receptieve vaardigheden (lezen, luisteren) die vaak objectief gecorrigeerd worden. Het bepalen van de cesuur is makkelijker voor de productieve vaardigheden. Lezen en luisteren zijn moeilijker omdat we mentale processen moeten interpreteren die alleen indirect te observeren zijn, waardoor een criterium-gerelateerd niveau moeilijk vast te leggen is.

Wanneer een toets verschillende vaardigheden omvat, moet een standaard bepaald worden voor elke afzonderlijke vaardigheid en bepaald worden hoe de resultaten samen te vatten (zie hoofdstuk 5.3 voor meer informatie hierover).

De lezer wordt uitgenodigd ook de *Manual for Relating Language Examinations to the CEFR* (Raad van Europa, 2009) te lezen. Die bevat een uitgebreidere beschrijving van de cesuurbepaling. Wat betreft de structuur en de terminologie van deze handleiding, geven we mee dat:

- hoofdstuk 6 over methodes voor cesuurbepaling enkel refereert aan objectief gecorrigeerde toetsen, dat wil zeggen voor lezen en luisteren;
- productieve vaardigheden worden behandeld onder de titel *Standardisation Training and Benchmarking* in hoofdstuk 5;
- hoofdstuk 7, dat gaat over VALIDERING, ook aandachtig zou moeten bekeken worden. Er bestaan twee algemene methodes van cesuurbepaling: taakgerichte en leerdergerichte. Taakgerichte methodes hangen af van het oordeel van deskundigen over items en dit is wat behandeld wordt in hoofdstuk 6. Bij leerdergerichte methodes wordt er gezocht naar aanvullend bewijs over de kandidaten dat de toets overstijgt. Dit wordt behandeld in hoofdstuk 7.
- de structuur niet impliceert dat taakgerichte cesuurbepaling belangrijker is dan leerdergerichte cesuurbepaling.

Strikt gesproken is een cesuurbepaling iets dat eenmalig zou moeten gebeuren, wanneer de toets voor het eerst wordt afgenomen. In de praktijk kan het bereiken van de beoogde standaard echter het resultaat zijn van een iteratieve procedure. Na verloop van tijd zouden we het waarderen van prestaties willen zien als een proces waarbij cesuren worden *bevestigd* en niet zozeer worden bepaald. Dit veronderstelt dat er geschikte procedures worden gehanteerd gedurende de volledige ontwikkelingscyclus. Dit wordt besproken in de aanvullende materialen die bij de handleiding horen (North en Jones, 2009).

## 5.3 Resultaten rapporteren

De gebruiker moet besluiten of enkel het toetsresultaat aan de kandidaat wordt gerapporteerd, of een resultatenprofiel die de prestatie op elk onderdeel van de toets weergeeft.

Het eerste is het meest gebruikelijke en geeft weer dat de meeste eindgebruikers van toetsresultaten een eenvoudig boven een complex antwoord lijken te verkiezen. De tweede geeft meer informatie die voor sommige doeleinden heel nuttig kan zijn.

Een derde mogelijkheid is dat beide worden gebruikt. Het ERK benadrukt dat het belangrijk is om profielscores te rapporteren waar het kan.

Waar een eenvoudig resultaat vereist wordt, moet er een methode gekozen worden om de score voor elk toetsonderdeel samen te voegen tot één enkele score. De gebruiker moet beslissen hoe elk toetsonderdeel wordt GEWOGEN: krijgen alle vaardigheden hetzelfde gewicht of wegen sommige zwaarder door? Dit kan enige aanpassing van de RUWE SCORE voor elk toetsonderdeel vereisen. Zie Bijlage VII.

Als de resultaten uitmonden in een certificaat, moet de gebruiker overwegen:

- Welk aanvullend materiaal (bv. 'can do'-descriptoren) verstrekt kan worden om de betekenis van het niveau te illustreren.
- Hoe de echtheid van het certificaat gegarandeerd kan worden (bv. falsificatie of veranderingen moeilijk maken, of een verificatiedienst voorzien).
- Welk voorbehoud (als dat er zou zijn) er bij de interpretatie van de resultaten moeten worden vermeld.

## 5.4 Kernvragen

- Hoeveel methodische correctie vereist uw toets en hoe vaak?
- Hoeveel beoordeling vereist uw toest en hoe vaak?
- Welk niveau van deskundigheid wordt er van uw beoordelaars verwacht?
- Hoe zal u garanderen dat de correctie en beoordeling accuraat en betrouwbaar is?
- Wat is de beste manier om in uw context kandidaten een score toe te kennen?
- Aan wie zal u de resultaten rapporteren en hoe zal u dit doen?

## 5.5 Aanbevolen literatuur

Zie ALTE (2006c) voor een checklist voor zelfevaluatie die gebruikt kan worden voor het toekennen van de scores, de waardering en de resultaten.

Kaftandjieva (2004), North en Jones (2009) en Figueras en Noijons (2009) geven informatie over cesuurbepalingen.

# 6 Monitoring en controle

Het is belangrijk om het werk dat is verricht om de toets te ontwikkelen en te gebruiken te controleren. Is het van een aannemelijke kwaliteit of zijn er verbeteringen nodig? Het doel van monitoren is te controleren of de belangrijke aspecten van een toets aanvaardbaar zijn tijdens of niet zo lang na het gebruik van de toets. Als er aanpassingen moeten gebeuren, dan is het vaak mogelijk om dit snel te doen. Verbeteringen kunnen in het voordeel zijn van de huidige kandidaten, of van diegenen die de volgende keer deelnemen aan de toets.

Toetsen controleren is het soort van project dat naar heel veel aspecten van de toets kijkt. Het gaat ook terug tot de toetsontwikkeling en stelt fundamentele vragen zoals 'is de toets nodig?', 'voor welk doel', 'voor wie?' en 'wat proberen we te toetsen?'. Het is zoals de fase van toetsontwikkeling, maar met dit voordeel dat er over data kan worden beschikt en over ervaring doordat de toets meerdere keren is gebruikt. Door hun omvang en doel kunnen toetscontroles geen onderdeel zijn van de normale toetscyclus en kunnen ze ook niet voor elke sessie uitgevoerd worden.

## 6.1 Routineobservatie

Monitoren is een onderdeel van de routineactiviteit van het samenstellen en gebruiken van een toets. Gegevens die tijdens routineobservaties worden verzameld, zullen gebruikt worden om ervoor te zorgen dat alles wat met de huidige versie van de toets te maken heeft, is zoals het hoort: materialen worden correct geproduceerd, ze worden op tijd afgeleverd, kandidaten krijgen de correcte scores toegewezen, enz. Daarna kunnen dezelfde gegevens gebruikt worden om in te schatten hoe het proces heeft gewerkt waarvoor werd gekozen, zoals het schrijven van de items en het vormgevingsproces, het SAMENSTELLEN VAN EEN TOETS, het toekennen van de scores, enz. De gegevens kunnen ook relevant zijn voor het VALIDITEITSBEWIJS (zie Bijlage I) en zouden ook gecontroleerd moeten worden met dit in het achterhoofd.

In deze handleiding werden al meerdere voorbeelden beschreven van hoe gegevens verzameld kunnen worden voor monitoring. Voorbeelden van monitoring omvatten:

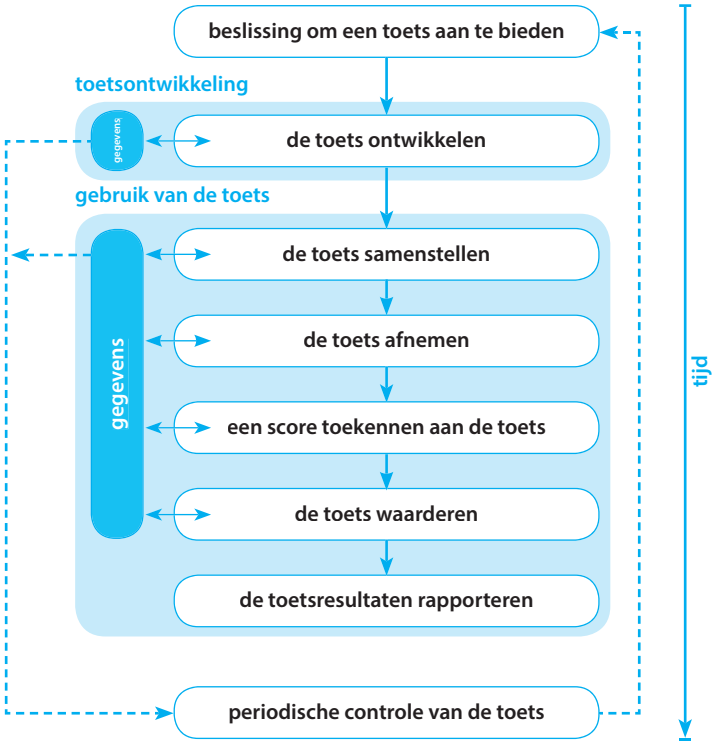
- het gebruik van het oordeel van deskundigen en PROEFAFNAMES of PRETESTEN om te garanderen dat alle items goed geschreven zijn (zie hoofdstuk 3.4);
- responsen van kandidaten gebruiken om te beslissen of items goed werken (zie Bijlage VII);
- feedbackformulieren gebruiken om te zien of de afname goed is verlopen (zie Bijlage VI);
- data verzamelen en analyseren over de prestaties van de correctoren (zie Bijlage VII).

De efficiëntie van het werk monitoren kan ook belangrijk zijn. Toetsaanbieders kunnen meten hoe lang een voorbereidende fase duurt en beslissen dat er te weinig of te veel tijd voor was voorzien.

## 6.2 Periodische controle van de toets

Periodische controle van de toets gebeuren occasioneel, buiten de normale werking van de toets. Dit kan na een bepaalde periode, op vaste tijdstippen, of na belangrijke veranderingen in de omstandigheden van de toets zoals wanneer de groep beoogde kandidaten verandert of het gebruik van de toets verandert, of de syllabus die gelinkt is aan de toets verandert. Het is ook mogelijk dat de nood aan controle naar voren kwam tijdens de monitoring. Controles laten toe om de toets en de manier waarop die werd samengesteld, uitgebreid onder de loep te nemen. Gegevens die verzameld werden tijdens het gebruik van de toets zoals bij het monitoren van de prestatie van correctoren, kunnen nuttig zijn voor een controle. De toetsaanbieder kan bovendien beslissen dat er andere gegevens nodig zijn en kan deze verzamelen voor de controle.

Tijdens een controle wordt er informatie verzameld en nauwkeurig geverifieerd. Deze informatie kan helpen om te beslissen hoe er met welke aspecten van de toets (bv. het construct, het format, de regels voor de afname) moet worden omgegaan. Het kan gebeuren dat de controle leidt tot kleine of helemaal geen veranderingen.




**Figuur 15** ***De toetscyclus en de periodische controle van de toets***

Figuur 15 is een reproductie van Figuur 5 (hoofdstuk 1.5.1) met als aanvulling de periodische controle. Het toont dat de bevindingen uit de controle aangewend worden in het allereerste stadium van het diagram: de beslissing om een toets aan te bieden. Ook het toetsontwikkelingsproces wordt opnieuw doorlopen als deel van de controle.

Het is belangrijk om veranderingen aan BELANGHEBBENDEN te melden, zoals gesuggereerd in hoofdstuk 2.6.

## 6.3 Aandachtspunten tijdens het monitoren en de controle

De monitoring en controle zijn onderdelen van de routine van de ontwikkeling van toetsen en hun gebruik. Ze vertellen de toetsaanbieder of alles werkt zoals het zou moeten en wat er zou moeten veranderen als dit niet het geval is. Controle kan ook helpen om anderen, zoals leidinggevenden in scholen of erkennende instellingen te tonen dat zij de toets kunnen vertrouwen. Als er wordt gekeken naar wat er is gebeurd en of het goed genoeg is gebeurd, dan gebeurt er eigenlijk een soort van audit van het VALIDITEITSBEWIJS.

ALTE (2007) heeft een lijst van 17 aandachtspunten opgesteld, de *Minimum Standards*, om toetsaanbieders te helpen hun validiteitsbewijs te structureren. Ze worden onderverdeeld in de volgende vijf algemene gebieden:

- samenstellen van de toets;
- afname en logistiek;
- score toekennen en waardering;
- analyse van de toets;
- communicatie met belanghebbenden.

Ze kunnen gebruikt worden met meer gedetailleerde en specifieke lijsten zoals de *ALTE Content Analysis Chekclists* (ALTE, 2004a-k, 2005, 2006a-c).

Er zijn ook ander tools beschikbaar die toetsaanbieders helpen om een validiteitsbewijs op te bouwen en te controleren. Jones, Smith en Talleu (2006, p. 490-2) geven een lijst met 31 aandachtspunten voor

voortgangstoetsen die door een beperkt aantal kandidaten worden afgelegd. Een groot deel van hun lijst is gebaseerd op de *Standards for Educational and Psychological Testing* (AERA et al., 1999).

## 6.4 Kernvragen

- Welke gegevens moeten er verzameld worden om de toets efficiënt te kunnen monitoren?
- Is er al een deel van deze gegevens verzameld om routinebeslissingen te nemen tijdens het gebruik van de toets? Hoe kunnen de gegevens makkelijk voor beide doeleinden gebruikt worden?
- Kunnen de gegevens bewaard worden en later bij de controle van de toets gebruikt worden?
- Wie moet er bij de controle betrokken worden?
- Welke middelen zijn er beschikbaar voor de controle?
- Hoe vaak zou de toets gecontroleerd moeten worden?
- Kan één van de lijsten met aandachtspunten gebruikt worden om het validiteitsbewijs te controleren?

## 6.5 Aanbevolen literatuur

ALTE (2007) biedt een aantal categorieën aan waarmee een toets kan geëvalueerd worden.

Zie ALTE (2002) voor een checklist voor zelfevaluatie die gebruikt kan worden voor de toetsanalyse en de controle.

Fulcher en Davidson (2009) illustreren een interessante manier om na te denken over het gebruik van gegevens tijdens de controle van een toets. Zij gebruiken de metafoor van een gebouw om te overwegen welke delen van een toets regelmatiger moeten veranderd worden en welke minder vaak moeten veranderen.

Beschrijvingen van verschillende aspecten van toetscontrole kunnen gevonden worden in Weir en Milanovic (2003).

# Bibliografie en hulpmiddelen


AERA, APA, NCME (1999) *Standards for Educational and Psychological Testing*, Washington DC: AERA Publishing.

Alderson, J C; Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

ALTE (1994) *Code of Practice*. Website. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/downloads/index.php**

ALTE (2002) *ALTE Quality Management and Code of Practice Checklist*: *test analysis and post examination review*. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/cop/copcheck.php**

ALTE (2004a) *Development and descriptive checklist for tasks and examinations*: *general*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004b) *Individual component checklist*: *reading*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004c) *Individual component checklist*: *structural competence*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004d) *Individual component checklist: listening*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004e) *Individual component checklist*: *writing*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004f) *Individual component checklist*: *speaking*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004g) *Individual component checklist – for use with one task*: *reading*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004h) *Individual component checklist – for use with one task*: *structural competence*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004i) *Individual component checklist – for use with one task*: *listening*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004j) *Individual component checklist – for use with one task*: *writing*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2004k) *Individual component checklist – for use with one task: speaking*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2005) *ALTE materials for the guidance of test item writers* (*1995, geüpdatet juli 2005*). Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2006a) *ALTE Quality Management and Code of Practice Checklist*: *test construction*. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/cop/copcheck.php**

ALTE (2006b) *ALTE Quality Management and Code of Practice Checklist*: *administration and logistics*. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/cop/copcheck.php**

ALTE (2006c) *ALTE Quality Management and Code of Practice Checklist: marking, grading, results*. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/cop/copcheck.php**

ALTE (2007) *Minimum standards for establishing quality profiles in ALTE Examinations*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2008a) *The ALTE Can Do Project*. Website. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

ALTE (2008b) *ALTE Quality Management and Code of Practice Checklists*. Website. Geraadpleegd op: 12/07/09. Beschikbaar op: **http://www.alte.org/cop/copcheck.php**

ALTE Members (1998) *Multilingual glossary of language testing terms* (Studies in Language Testing volume 6), Cambridge: Cambridge University Press.

ALTE Members (2005a) *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20INput51.pdf**

ALTE Members (2005b) *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20OUTput51.pdf**

ALTE Members (2007a) *The CEFR Grid for Writing Tasks v. 3.1 (analyse)*. Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_1_analysis.doc**

ALTE Members (2007b) *The CEFR Grid for Writing Tasks v. 3.1 (presentatie)*. Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_1_presentation.doc**

ALTE Working Group on Code of Practice (2001) *The Principles of Good Practice for ALTE Examinations*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/downloads/index.php**

Assessment Systems (2009) Iteman 4. Software. Assessment Systems.

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press. Bachman, L F (2004) *Statistical Analysis for Language Assessment*, Cambridge: Cambridge University Press.

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly* 2 (1), 1–34.

Bachman, L F; Black, P; Frederiksen, J; Gelman, A; Glas, C A W; Hunt, E; McNamara, T and Wagner, R K (2003) Commentaries Constructing an Assessment Use Argument and Supporting Claims About Test Taker-Assessment Task Interactions in Evidence-Centered Assessment Design, *Measurement: Interdisciplinary Research & Perspective* 1 (1), 63–91.

Bachman, L F and Palmer, A S (1996) *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L F and Palmer, A S (2010) *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Banerjee, J (2004) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section D: Qualitative Analysis Methods*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionD.pdf**

Beacco J-C and Porquier, R (2007) *Niveau A1 pour le français. Un référentiel*, Paris: Editions Didier.

Beacco J-C and Porquier, R (2008) *Niveau A2 pour le français. Un référentiel*, Paris: Editions Didier.

Beacco, J-C; Bouquet, S and Porquier, R (2004) *Niveau B2 pour le français. Un référentiel*, Paris: Editions Didier.

Bolton, S; Glaboniat, M; Lorenz, H; Perlmann-Balme, M and Steiner, S (2008) *Mündliche – Mündliche Produktion und Interaktion Deutsch: Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*, München: Langenscheidt.

Bond, T G and Fox, C M (2007) *Applying the Rasch model: fundamental measurement in the human sciences*, Mahwah, NJ: Lawrence Erlbaum.

Brennan, R L (1992) Generalizability Theory, *Instructional Topics in Educational Measurement Series 14*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/21.pdf**

Briggs, D C; Haertel, E; Schilling, S G; Marcoulides, G A and Mislevy, R J (2004) Comment: Making an Argument for Design Validity Before Interpretive Validity, *Measurement: Interdisciplinary Research & Perspective* 2 (3), 171–191.

Camilli, G and Shepard, L A (1994) *Methods for Identifying Biased Test Items*, Thousand Oaks, CA: Sage.

Canale, M and Swain, M (1981) A theoretical framework for communicative competence, in Palmer, A S; Groot, P J and Trosper, S A (Red.) *The Construct Validation of Tests of Communicative Competence*, Washington DC: TESOL.

Carr, N T (2008) Using Microsoft Excel® to Calculate Descriptive Statistics and Create Graphs, *Language Assessment Quarterly* 5 (1), 43.

CEFTrain (2005) *CEFTrain*. Website. Geraadpleegd op: 04/03/09.
Beschikbaar op: **http://helsinki.fi/project/ceftrain/index.html**

Chapelle, C A; Enright, M K and Jamieson, J M (2007) *Building a Validity argument for the Test of English as a Foreign Language*, Oxford: Routledge.

CIEP (2009) *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues*. Website. Geraadpleegd op: 12/07/09. Beschikbaar op: **www.ciep.fr/publi_evalcert**

CIEP/Eurocentres (2005) *Exemples de productions orales illustrant, pour le français, les niveaux du Cadre européen commun de référence pour les langues*, DVD, Strasbourg: Council of Europe.

Cizek, G J (1996) Standard-setting guidelines, *Instructional Topics in Educational Measurement Series*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/Standard.pdf**

Cizek, G J and Bunch, M B (2006) *Standard Setting: A Guide To Establishing And Evaluating Performance Standards On Tests*, Thousand Oaks, CA: Sage.

Cizek, G J; Bunch, M B and Koons, H (2004) *Setting performance standards: contemporary methods, Instructional Topics in Educational Measurement Series*. Geraadpleegd op: 05/03/09. Gedownload van:
**http://www.ncme.org/pubs/items/Setting%20Performance%20Standards%20ITEMS%20Module.pdf**

Clauser, B E and Mazor, K M (1998) Using statistical procedures to identify differentially functioning test items, *Instructional Topics in Educational Measurement Series*. Geraadpleegd op: 05/03/09. Gedownload van:
**http://www.ncme.org/pubs/items/Statistical.pdf**

Cook, L L and Eignor, D R (1991) IRT Equating Methods, *Instructional Topics in Educational Measurement Series 10*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/17.pdf**

Corrigan, M (2007) *Seminar to calibrate examples of spoken performance, Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica) Perugia, 17th–18th December 2005*. Geraadpleegd op: 07/03/10. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/Report_Seminar_Perugia05.pdf**

Coste, D (2007) *Contextualising Uses of the Common European Framework of Reference for Languages*, paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg 2007; Gedownload van:
**http://www.coe.int/t/dg4/linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc**

Council of Europe (1996) *Users' Guide for Examiners*, Strasbourg: Language Policy Division.

Council of Europe (1998) *Modern Languages: learning, teaching, assessment. A Common European Framework of Reference*, Strasbourg: Language Policy Division.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Council of Europe (2004a) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Geraadpleegd op: 04/03/09. Gedownload van:
**http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf**

Council of Europe (2005) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French, German, Italian and Spanish, CD, Strasbourg: Council of Europe.

Council of Europe (2006a) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Geschreven voorbeelden. Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/T/DG4/Portfolio/documents/exampleswriting.pdf**

Council of Europe (2006b) *TestDaF Sample Test Tasks*. Geraadpleegd op: 04/03/09. Gedownload van: **http://www. coe.int/T/DG4/Portfolio/documents/ALTECEFR%20Writing%20Grid-2.0_TestDaF%20samples.pdf**

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A Manual*. Geraadpleegd op: 04/03/09. Gedownload van:
**http://www.coe.int/t/dg4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf**

Council of Europe and CIEP (2009) *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues*, DVD, Strasbourg and Sèvres: Council of Europe and CIEP.

Davidson, F and Lynch, B K (2002) *Testcraft*: *A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

Davies, A (1997) (Guest Ed.) Ethics in language testing, *Language Testing* 14 (3).

Davies, A (2004) (Guest Ed.) *Language Assessment Quarterly* 2 & 3.

Davies, A (2010) Test fairness: a response, *Language Testing* 27 (2), 171-176.

Davies, A; Brown, A; Elder, C; Hill, K; Lumley, T and McNamara, T (1999) *Dictionary of language testing* (Studies in Language Testing volume 7), Cambridge: Cambridge University Press.

Downing, S M and Haladyna, T M (2006) *Handbook of Test Development*, Mahwah, NJ: Lawrence Erlbaum.

EALTA (2006) *EALTA Guidelines for Good Practice in Language Testing and Assessment*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ealta.eu.org/documents/archive/guidelines/English.pdf**

Eckes, T (2009) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF*: *section H*: Many-Facet Rasch Measurement. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf**

Education Testing Services (2002) *ETS Standards for Quality and Fairness*, Princeton, NJ: ETS.

Embretson, S E (2007) Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure?,

*Educational Researcher* 36 (8), 449.

Eurocentres and Federation of Migros Cooperatives (2004) *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Languages*, DVD, Strasbourg: Council of Europe.

Europarat; Council for Cultural Co-operation, Education Committee, Modern Languages Division; Goethe-Institut Inter Nationes u.a. (Hg.) (2001) *Gemeinsamer europäischer Referenzrahmen für Sprachen*: *lernen*, *lehren*, *beurteilen*, Berlin, München: Langenscheidt.

Figueras, N; Kuijper, H; Tardieu, C; Nold, G and Takala, S (2005) *The Dutch Grid Reading/Listening*. Website. Geraadpleegd op: 04/03/09. Beschikbaar op: **http://www.lancs.ac.uk/fss/projects/grid/**

Figueras, N and Noijons, J (Red.) (2009) *Linking to the CEFR levels*: *Research perspectives*. CITO/Council of Europe. Geraadpleegd op: 10/01/10. Gedownload van: **http://www.coe.int/t/dg4/linguistic/EALTA_PublicatieColloquium2009.pdf**

Frisbie, D A (1988) Reliability of Scores From Teacher-Made Tests, *Instructional Topics in Educational Measurement Series 3*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/ITEMS_Mod_3.pdf**

Fulcher, G and Davidson, F (2007) *Language Testing and Assessment – an advanced resource book*, Abingdon: Routledge.

Fulcher, G and Davidson, F (2009) Test architecture, test retrofit, *Language Testing* 26 (1), 123–144.

Glaboniat, M; Müller, M; Rusch, P; Schmitz, H and Wertenschlag, L (2005) *Profile Deutsch – Gemeinsamer europäischer Referenzrahmen*. *Lernzielbestimmungen*, *Kannbeschreibungen*, *Kommunikative Mittel*, *Niveau A1–A2*, *B1– B2*, *C1–C2*, Berlin, München: Langenscheidt.

Goethe-Institut Inter Nationes; der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK); der Schweizerischen Konferenz der Kantonalen Erziehungsdirektoren (EDK) und dem österreichischen Bundesministerium für Bildung, Wissenschaft und Kultur (BMBWK), (Hg.) (2001) *Gemeinsamer Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*, im Auftrag des Europarats, Rat für kulturelle Zusammenarbeit, deutsche Ausgabe, München: Langenscheidt.

Gorin, J S (2007) Reconsidering Issues in Validity Theory, *Educational Researcher* 36 (8), 456.

Grego Bolli, G (Red.) (2008) *Esempi di Produzioni Orali – A illustrazione per l'italiano dei livelli del Quadro comune europeo di riferimento per le lingue*, DVD, Perugia: Guerra Edizioni.

Haertel, E H (1999) Validity arguments for High-Stakes Testing: in search of the evidence, *Educational Measurement*: *Issues and Practice* 18 (4), 5.

Hambleton, R K and Jones, R W (1993) Comparison of classical test theory and item response theory and their applications to test development, *Instructional Topics in Educational Measurement Series 16*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/24.pdf**

Harvill L M (1991) Standard error of measurement, *Instructional Topics in Educational Measurement Series 9*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/16.pdf**

Heaton, J B (1990) *Classroom Testing*, Harlow: Longman.

Holland, P W and Dorans, N J (2006) Linking and Equating, in Brennan, R L (Red.) *Educational measurement* (*4e editie*), Washington, DC: American Council on Education/Praeger.

Hughes, A (1989) *Testing for Language Teachers*, Cambridge: Cambridge University Press.

ILTA (2000) *ILTA Code of Ethics*. Website. Geraadpleegd op: 05/03/09. Beschikbaar op: **http://www.iltaonline.com/index.php?option=com_content&task=view&id=57&Itemid=47**

ILTA (2007) *ILTA Guidelines for Practice*. Website. Geraadpleegd op: 05/03/09. Beschikbaar op: **http://iltaonline.com/index.php?option=com_content&task=view&id=122&Itemid=133**

Instituto Cervantes (2007) *Plan curricular del Instituto Cervantes – Niveles de referencia para el español*, Madrid: Edelsa.

JCTP (1988) *Code of Fair Testing Practices in Education*. Geraadpleegd op: 12/08/09. Gedownload van: **http://www.apa.org/science/fairtestcode.html**

JLTA (not specified) *Code of Good Testing Practice*. Geraadpleegd op: 12/08/09. Gedownload van: **http://www.avis.ne.jp/~youichi/COP.html**

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning and the CEFR, *Annual Review of Applied Linguistics* 29, 51–63.

Jones, P; Smith, R W and Talley, D (2006) Developing Test Forms for Small-Scale Achievement Testing Systems, in Downing, S M and Haladyna, T M (Red.) *Handbook of Test Development*, Mahwah, NJ: Lawrence Erlbaum.

Jones, R L and Tschirner, E (2006) *A Frequency Dictionary of German – Core Vocabulary for Learners*, New York: Routledge.

Kaftandjieva, F (2004) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF*: *section B*: *Standard Setting*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf**

Kane, M (2002) Validating High Stakes Testing Programs, *Educational Measurement*: *Issues and Practices* 21 (1), 31–41.

Kane, M (2004) Certification Testing as an Illustration of Argument-Based Validation, *Measurement*: *Interdisciplinary Research & Perspective* 2 (3), 135–170.

Kane, M (2006) Validation, in Brennan, R L (Red.) *Educational measurement* (*4th edition*), Washington, DC: American Council on Education/Praeger.

Kane, M (2010) Validity and fairness, *Language Testing* 27(2), 177-182.

Kane, M; Crooks, T and Cohen, A (1999) Validating measures of performance, *Educational Measurement*: *Issues and Practice* 18 (2), 5–17.

Kolen, M J (1988) Traditional Equating Methodology, *Instructional Topics in Educational Measurement Series 6*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/11.pdf**

Kolen, M J (2006) Scaling and Norming, in Brennan, R L (Red.) *Educational measurement* (*4e editie*), Washington, DC: American Council on Education/Praeger.

Kuijper, H (2003) *QMS as a Continuous Process of Self-Evaluation and Quality Improvement for Testing Bodies*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/qa/index.php**

Kunnan, A J (2000a) Fairness and justice for all, in Kunnan, A J (Red.) *Fairness and validation in language assessme*nt , Cambridge: Cambridge University Press, 1–13.

Kunnan, A J (2000b) *Fairness and Validation in Language Assessment*: *Selected papers from the 19th Language Testing Research Colloquium*, *Orlando*, *Florida* (Studies in Language Testing volume 9), Cambridge: Cambridge University Press.

Kunnan, A J (2004) Test Fairness, in Milanovic, M and Weir, C (Eds) *European Language Testing in a Global Context – Proceedings of the ALTE Barcelona Conference*, *July 2001* (Studies in Language Testing volume 18), Cambridge: Cambridge University Press.

Linacre, J M (2009) Facets 3.64.0. Software. Winsteps.com software.

Lissitz, R W and Samuelsen, K (2007a) A Suggested Change in Terminology and Emphasis Regarding Validity and Education, *Educational Researcher* 36 (8), 437.

Lissitz, R W and Samuelsen, K (2007b) Further Clarification Regarding Validity and Education, *Educational Researcher* 36 (8), 482.

Livingston, S (2004) *Equating Test Scores* (*Without IRT*). Geraadpleegd op: 12/07/09. Gedownload van: **http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf**

McNamara, T and Roever, C (2006) Fairness Reviews and Codes of Ethics, *Language Learning* 56 (S2), 129–148.

Messick, S (1989) Meaning and values in test validation: the science and ethics of assessment, *Educational Researcher: Issues and Practice* 18, 5–11.

Messick, S (1989) Validity, in Linn, R (Ed.) *Educational measurement*, 3rd edition, New York: Macmillan, 13–103. Mislevy, R J (2007) Validity by Design, *Educational Researcher* 36 (8), 463.

Mislevy, R J; Steinberg, L S and Almond, R G (2003) Focus Article: On the Structure of Educational Assessments, *Measurement*: *Interdisciplinary Research & Perspective* 1 (1), 3–62.

Moss, P A (2007) Reconstructing Validity, *Educational Researcher* 36 (8), 470.

Nederlandse Taalunie (2008). *Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen: Leren, Onderwijzen, Beoordelen.* Den Haag: Taalunie.

North, B and Jones, N (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages*: *Learning*, *teaching*, *assessment* (*CEFR*) – *Further Material on Maintaining Standards across Languages*, *Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*. Geraadpleegd op: 04/03/09. Gedownload van: **http://www.coe.int/t/dg4/linguistic/Manual%20-%20Extra%20Material%20-%20 proofread%20-%20FINAL.pdf**

Parkes, J (2007) Reliability as Argument, *Educational Measurement*: *Issues and Practice* 26(4): 2-10.

Perlmann-Balme, M and Kiefer, P (2004) *Start Deutsch*. *Deutschprüfungen für Erwachsene*. *Prüfungsziele*, *Testbeschreibung*, München, Frankfurt: Goethe-Institut und WBT.

Perlmann-Balme, M; Plassmann, S and Zeidler, B (2009) *Deutsch-Test für Zuwanderer*. *Prüfungsziele*, *Testbeschreibung*, Berlin: Cornelsen.

Saville, N (2005) Setting and monitoring professional standards: A QMS approach, *Research Notes* 22. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.cambridgeesol.org/rs_notes/rs_nts22.pdf**

Sireci, S G (2007) On Validity Theory and Test Validation, *Educational Researcher* 36 (8), 477.

Spinelli, B and Parizzi, F (2010) *Profilo della lingua italiana*. *Livelli di riferimento del QCER A1, A2, B1 e B2,* Milan: RCS libri – Divisione education.

Spolsky, B (1981) Some ethical questions about language testing, in Klein-Braley, C and Stevenson, D (Red.) *Practice and problems in language testing*, Frankfurt: Verlag Peter Lang, 5–21.

Stiggins, R J (1987) Design and Development of Performance Assessment, *Instructional Topics in Educational Measurement Series 1*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/ITEMS_Mod_1_Intro.pdf**

Traub, R E and Rowley, G L (1991) Understanding reliability, *Instructional Topics in Educational Measurement Series 8*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/15.pdf**

Trim, J L M (2010) Plenary presentation at ACTFL-CEFR Alignment Conference, Leipzig, Juni 2010.

University of Cambridge ESOL Examinations (2004) *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Language*, DVD, Strasbourg: Council of Europe.

University of Cambridge ESOL Examinations/Council of Europe (2009a) *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2*, DVD, Cambridge: University of Cambridge ESOL Examinations.

University of Cambridge ESOL Examinations/Council of Europe (2009b) *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2*. Website. Geraadpleegd op: 02/09/09. Beschikbaar op: **http://www.cambridgeesol.org/what-we-do/research/speaking-performances.html**

van Avermaet, P (2003) *QMS and The Setting of Minimum Standards: Issues of Contextualisation Variation between The Testing Bodies*. Geraadpleegd op: 12/07/09. Gedownload van: **http://www.alte.org/qa/index.php**

van Avermaet, P; Kuijper, H and Saville, N (2004) A Code of Practice and Quality Management System for International Language Examinations, *Language Assessment Quarterly* 1 (2 & 3), 137–150.

van Ek, J A and Trim, J (1990) *Waystage 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J (1991) *Threshold 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J (2001) *Vantage*, Cambridge: Cambridge University Press.

Verhelst, N (2004a) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section C: Classical Test Theory*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionC.pdf**

Verhelst, N (2004b) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section E: Generalizability Theory*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionE.pdf**

Verhelst, N (2004c) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section F: Factor Analysis*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionF.pdf**

Verhelst, N (2004d) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section G: Item Response Theory*. Geraadpleegd op: 09/06/10. Beschikbaar op: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionG.pdf**

Ward, A W and Murray-Ward, M (1994) Guidelines for Development of item banks, *Instructional Topics in Educational Measurement Series 17*. Geraadpleegd op: 05/03/09. Gedownload van: **http://www.ncme.org/pubs/items/25.pdf**

Weir, C J (2005) *Language testing and validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Weir, C J and Milanovic, M (2003) (Red.) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (Studies in Language Testing volume 15), Cambridge: Cambridge University Press.

Widdowson, H G (1978) *Language Teaching as Communication*, Oxford: Oxford University Press.

Xi, X (2010) How do we go about investigating test fairness?, *Language Testing* 27(2): 147-170.

## Websites

Association of Language Testers in Europe: **www.alte.org**

English Profile: **www.englishprofile.org**

European Association for Language Testing and Assessment: **http://www.ealta.eu.org/**

International Language Testing Association: **http://www.iltaonline.com/**

Language Policy Division, Council of Europe: **http://www.coe.int/t/dg4/linguistic/default_EN.asp**

# Bijlage I – Een validiteitsbewijs opbouwen

Deze bijlage introduceert een benadering van VALIDERING die de opbouw van een VALIDITEITSBEWIJS bevat. Het is meer gedetailleerd dan het overzicht in 1.2.3 en toont hoe de stappen in de argumentatie eigenlijk niet afgebakend en sequentieel zijn, maar overlappend en met elkaar verbonden.

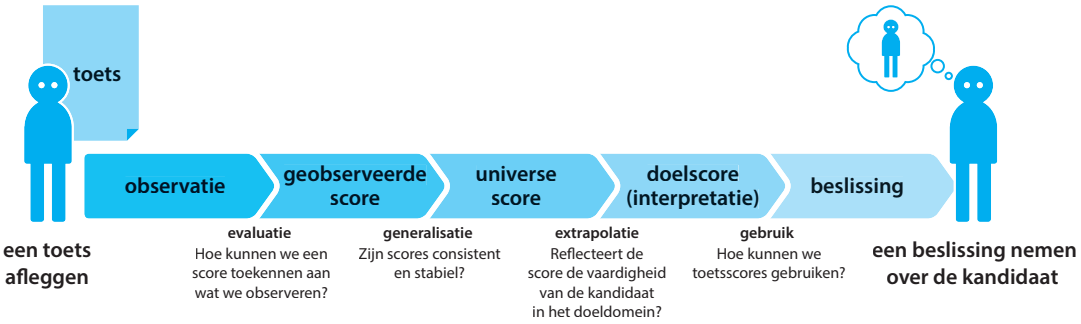
Kane (2006), Kane, Crookes en Cohen (1999), Bachman (2005) en Bahman en Palmer (2010) beschrijven validiteitsbewijzen uitgebreider. Validering is een voortdurend proces, waarbij aan het validiteitsbewijs alsmaar meer gegevens worden toegevoegd waardoor het alsmaar genuanceerder onderbouwd wordt.

De focus van een validiteitsbewijs is de interpretatie en het gebruik van de toetsresultaten en daarbij wordt deze definitie van validiteit gevolgd: '*the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests*' (AERA et al., 1999).

Een validiteitsbewijs is een reeks van beweringen die beschrijven waarom voorgestelde interpretaties van toetsresultaten valide zijn, en biedt de gegevens en de theorie om dit te onderbouwen. Deze bijlage geeft een overzicht van hoe dit kan worden gedaan.

Als het validiteitsbewijs gepresenteerd wordt aan BELANGHEBBENDEN, dan is het vertrekpunt een duidelijke uiteenzetting van de manier waarop de toetsresultaten moeten geïnterpreteerd worden voor een bepaald gebruik. Het GEBRUIKSBEWIJS (ook wel het INTERPRETATIEF BEWIJS genoemd) legt dit uit. Het validiteitsbewijs verantwoordt het interpretatief bewijs theoretisch en empirisch.

Figuur 16 toont een conceptuele voorstelling van een gebruiksbewijs naar het voorbeeld van Bachman (2005). Het is een ketting van redeneringen, in vier stappen (elke stap wordt door een pijl afgebeeld), die het gebruik van de toetsresultaten verantwoordt. Elke stap geeft een conceptuele basis voor de stap die volgt. Betrouwbare toetsresultaten bijvoorbeeld (*universe score*) zijn alleen bruikbaar als ze adequaat een prestatie op een toets weergeven (geobserveerde score). Het diagram stelt *geen* opeenvolging van fases voor die achtereenvolgens moeten worden afgewerkt. De gegevens om elke stap te verantwoorden kunnen worden vergaard tijdens verschillende fases van het toetsontwikkelingsproces en het samenstellen van een toets.



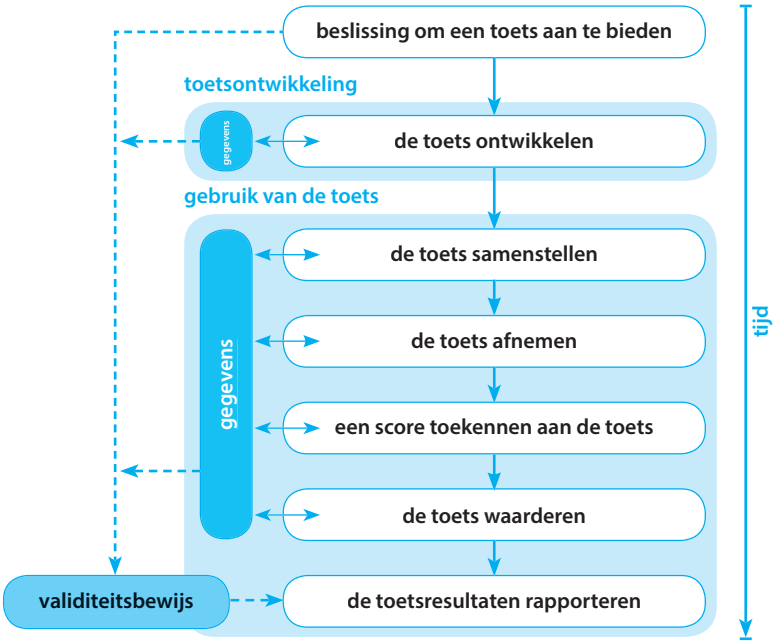

**Figuur 16 *Ketting van redeneringen in een validiteitsbewijs* (gebaseerd op Kane, Crooks, Cohen 1999, Bachman 2005)**

Het validiteitsbewijs wordt opgebouwd om het gebruiksbewijs te ondersteunen en bestaat uit gegevens, theorie en beargumenteerde beweringen. De gegevens die elke stap onderbouwen, worden tijdens de toetsontwikkeling, het samenstellen van een toets en het toetsgebruik verzameld.

Veel van de gegevens die deel uitmaken van het validiteitsbewijs resulteren uit het routineuze gebruik van de toets. Voorbeelden hiervan worden opgelijst in 6.1. De gegevens die voor een ander, meer direct doel verzameld worden, zoals de prestatie van beoordelaars monitoren, zal ook nuttig zijn om een validiteitsbewijs op te bouwen. Dit wordt geïllustreerd in Figuur 17.

Het validiteitsbewijs kan verbeterd en ontwikkeld worden door gegevens te verzamelen telkens er een nieuwe vorm van de toets ontwikkeld en gebruikt wordt. De ontwikkeling van het validiteitsbewijs zou moeten beginnen in de vroegste fase van het proces, wanneer de beoogde doelen van de toets vastgelegd worden. Veel van het validiteitsbewijs van een bepaalde toets kan echter ook gebruikt worden voor het validiteitsbewijs van een nieuwe vorm van de toets.

Sommige theoretici (Bachman, 2005; Mislevy et al., 2003) benadrukken dat een validiteitsbewijs de vorm zou moeten aannemen van een *informeel bewijs*. Dit betekent dat het bewijs door logisch redeneren alleen niet als juist of fout bevonden kan worden. Het kan echter door iemand die het nagaat meer of minder overtuigend bevonden worden. Hoe overtuigend het is, zal afhangen van de theorie en de gegevens die beschikbaar zijn om het te onderbouwen.




**Figuur 17** ***De toetscyclus, periodische controle en het validiteitsbewijs***

Het validiteitsbewijs zou minder overtuigend kunnen worden door nieuwe gegevens of een nieuwe theorie, of wanneer bestaande gegevens anders worden geïnterpreteerd. De kans bestaat ook dat toetsaanbieders, ongewild, eenvoudigweg de interpretaties bevestigen die zij verkiezen, zonder kritisch genoeg te zijn. Nadat het validiteitsbewijs voor de eerste keer werd geleverd, moet het daarom blijvend in vraag worden gesteld, ook al betekent dit dat de aanbevolen interpretaties van de toetsresultaten daardoor kunnen veranderen. Dit zou bijvoorbeeld kunnen worden gedaan door alternatieve mogelijkheden om de gegevens te begrijpen te overwegen of door na te gaan of alle conclusies die het bewijs levert voldoende solide zijn. De toetsaanbieders zouden dan hun validiteitsbewijs moeten herbekijken, de nodige aanpassingen doen en argumenteren waarom de gegevens op een bepaalde manier geïnterpreteerd werden.

Voorbeelden van gegevens die kunnen gebruikt worden om het validiteitsbewijs te onderbouwen, worden in de rest van deze bijlage gegeven. Ook worden er voorbeelden aangereikt van verschillende manieren om de gegevens te begrijpen. Al deze voorbeelden zijn gebaseerd op het werk van Kane (2004) en Bachman (2005). Ze zijn volgens de structuur van deze handleiding geordend: toetsen ontwikkelen, toetsen samenstellen, toetsen afnemen, de toets een score toekennen en waarderen en de resultaten rapporteren. Toetsaanbieders kunnen deze gegevens als een vertrekpunt zien voor de samenstelling van hun eigen validiteitsbewijzen; de lijsten zijn echter niet exhaustief.

**Aanbevolen literatuur**

ALTE (2005:19) geeft een nuttig overzicht van soorten validiteit en de achtergrond van de moderne concepten van validiteit.

AERA et al (1999) geeft een overzicht van de moderne concepten van validiteit en standaarden die specifieke aandachtspunten onderstrepen en kan daarom een hulp zijn bij de ontwikkeling van een validiteitsbewijs.

Messick (1989) behandelt het unitaire concept van validiteit en ook de ethische overwegingen die ermee samenhangen.

Haertel (1999) geeft een voorbeeld van hoe gegevens en argumentatie verbonden kunnen zijn met de interpretatie van scores.

Kane, Crooks en Cohen (1999) presenteren een duidelijk beeld van de eerste stappen in het opbouwen van een validiteitsbewijs. Kane (2006) geeft een diepgaandere uiteenzetting.

Bachman (2005) bespreekt validiteitsbewijzen in relatie tot taaltoetsing. Hij bespreekt ook het model dat Bachman en Palmer (1996) voorstelden als model van validiteitsbewijs. In het eerdere model werd het nut gezien als de belangrijkste eigenschap van een toets aangezien het nut het evenwicht houdt tussen betrouwbaarheid, validiteit, authenticiteit, interactiviteit en impact.

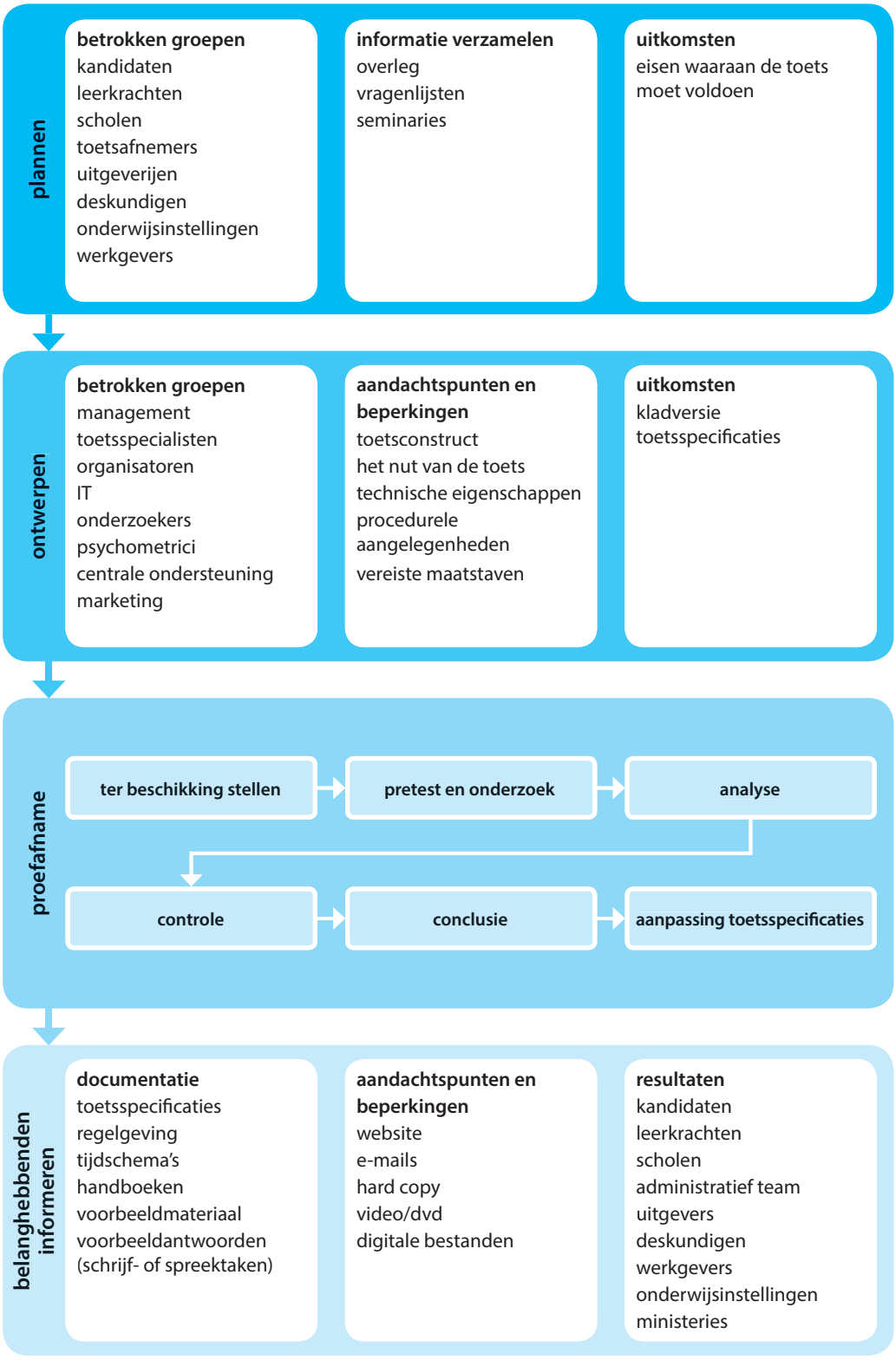
Bachman en Palmer (2010) tonen aan hoe validiteitsbewijzen fundamenteel zijn bij toetsontwikkeling en een referentiekader kunnen bieden voor deze taken.

| | Evaluatie | Generalisatie | Extrapolatie | Gebruik |
|---|---|---|---|---|
| | **Hoe kunnen we een score toekennen aan wat we observeren?** | **Zijn scores consistent en stabiel ?** | **Reflecteert de score de vaardigheid van de kandidaat in het doeldomein?** | **Hoe kunnen we de toetsscores gebruiken?** |
| **EEN TOETS ONTWIKKELEN** (hoofdstuk 2) | | | | |
| **Bewijs voor** | | De toetsspecificaties vereisen een standaard toetsformat. Dit zal het idee ondersteunen dat verschillende versies van toetsen op elkaar lijken (zie hoofdstuk 2 en Bijlage III). | De richtlijnen voor de itemschrijvers en de toetsspecificaties zijn duidelijk bedoeld voor een welbepaald gebruiksdomein. Dit domein kan ook geïdentificeerd worden in een behoefteonderzoek (zie hoofdstuk 2.4).<br>Bewijs dat de cesuur juist is bepaald zal de aanbevolen interpretatie van de toetsresultaten van elke kandidaat ondersteunen (zie hoofdstuk 2.0 en 5.2). | |
| **Bewijs tegen** | | | Sommige aspecten van het CONSTRUCT zijn niet volledig vertegenwoordigd in de toetsspecificaties. Dit zou betekenen dat de toetsresultaten niet op een adequate manier informatie verstrekken over wat de kandidaten kunnen doen (zie hoofdstuk 1.1 en 2). | |

| | Evaluatie | Generalisatie | Extrapolatie | Gebruik |
|---|---|---|---|---|
| **EEN TOETS SAMENSTELLEN** (hoofdstuk 3) | | | | |
| **Bewijs voor** | Alle correctiesleutels zijn correct.<br>Grammaticaboeken, woordenboeken en kennis van deskundigen helpen dit na te gaan. | Items in een bepaalde toestvorm representeren het construct even goed als de items in een andere toestvorm. Het is onwaarschijnlijk dat elke keer exact dezelfde elementen aan bod komen, maar de elementen van het construct moeten op een vergelijkbare manier geselecteerd worden (zie hoofdstuk 2, 3.5 en Bijlage VII).<br>De manier waarop toetsvormen worden gerelateerd is geschikt (zie Bijlage VII).<br>Als er statistische analyses werden gebruikt, dan werd er gevonden dat de meetfout klein is en de statistische modellen aansluiten bij de data (MODEL FIT, zie Bijlage VII). | Er werden deskundigen ingeschakeld bij het schrijven van de items en het samenstellen van een toets (zie hoofdstuk 3.2). | |
| **Bewijs tegen** | | Verschillende versies van een toets zijn niet met elkaar gelinkt.<br>Verschillende versies van een toets representeren niet hetzelfde construct. | Het is mogelijk dat sommige elementen van het construct niet helemaal adequaat werden gerepresenteerd in de toetsmaterialen. Dit zou betekenen dat de resultaten geen adequate informatie geven over wat de kandidaten kunnen. | |
| **EEN TOETS AFNEMEN** (hoofdstuk 4) | | | | |
| **Bewijs voor** | Er werden tijdens de afname toetsprocedures gevolgd. Dit zal helpen om aan te tonen dat de toetsscore niet beïnvloed wordt door andere factoren (zoals te veel of te weinig tijd). (zie hoofdstuk 4.2). | De procedures werden altijd gevolgd en helpen aantonen dat de verschillende versies van de toets over verschillende afnames vergelijkbaar zijn. (zie hoofdstuk 4.2). | | |
| **Bewijs tegen** | Als er geen gevolgd wordt gegeven aan afkijken, zal de toesscore de vaardigheid van de kandidaat niet weerspiegelen. | Onopgemerkt spieken zal betekenen dat de scores van sommige kandidaten hun taalvaardigheid onvoldoende weerspiegelen. Dit zal waarschijnlijk verschillend zijn over de versies van de toets heen. | Externe factoren kunnen de toetsresultaten beïnvloed hebben. Dit zou kunnen doordat bv. de afnameprocedures niet gevolgd werden. Het gevolg is dat toetsresultaten ook deze externe factoren weerspiegelen en het moeilijk is om te zeggen dat ze enkel gerelateerd zijn aan taalvaardigheid (zie hoofdstuk 4). | |

| | Evaluatie | Generalisatie | Extrapolatie | Gebruik |
|---|---|---|---|---|
| **SCORES TOEKENNEN, WAARDEREN EN RESULTATEN RAPPORTEREN** (hoofdstuk 5) | | | | |
| **Bewijs voor** | Tijdens het toekennen van de scores werden de procedures gevolgd. Dit zal helpen aantonen dat de toetsscore niet beïnvloed werd door andere factoren (zoals het gebruik van een foute correctiesleutel, of fouten bij het scannen) (zie hoofdstuk 5.0).<br>De correctie was accuraat en betrouwbaar (zie hoofdstuk 5.1 en Bijlage VII). | Bewijs van de betrouwbaarheid van de score (meestal statistisch bewijs), kan tonen dat deze versie van de toets de kandidaten op een consistente manier meet (zie hoofdstuk 1.3, 5.1 en Bijlage VII).<br>Als enkel de data van een beperkt aantal kandidaten worden geanalyseerd, zijn deze data representatief voor de hele groep kandidaten (zie Bijlage VII).<br>Als de cesuren lage niveaus van meetfout hebben, dan betekent dit dat de kandidaten waarschijnlijk correct gewaardeerd worden (zie Bijlage VII). | Het gebruik van deskundige beoordelaars impliceert dat de beoordelingen waarschijnlijk het vooropgestelde domein meer weerspiegelen (zie hoofdstuk 5.1).<br>Vergelijkbaar hiermee is het gebruik van goed geschreven beoordelingsschalen die ervoor zorgen dat de prestaties nog meer beoordeeld worden in overeenstemming met het doeldomein (zie hoofdstuk 2.5 en 5.1.3). | Als er regels worden gebruikt om specifieke beslissingen te nemen, gebaseerd op de toetsresultaten, dan is het waarschijnlijk dat de toets zal gebruikt worden zoals gepland en dat ongewenste effecten tot het minimum zullen worden beperkt. (zie hoofdstukken 1.2, 5.3 en Bijlage I). |
| **Bewijs tegen** | | Als er voor de analyses data van een niet-representatieve groep van kandidaten werden gebruikt, dan kunnen die fouten en/of bias bevatten (zie Bijlage VII). | | Als de regels en standaardprocedures niet gevolgd werden bij het nemen van beslissingen, dan kan de toets onjuist gebruikt worden (zie hoofstukken 1.5 en 5.3). |

# Bijlage II – Het toetsontwikkelingsproces



# Bijlage III – Voorbeeld van een toetsformat[4]

## Inhoud en overzicht

| Toetsformulier/ Timing | Format | | Aantal vragen | Focus van de toets |
|---|---|---|---|---|
| **LEZEN**<br>1 uur | **Deel 1** | Een MATCHING-taak met een doorlopende tekst die in vier delen of vier informatieve teksten is verdeeld; ongeveer 205-350 woorden in totaal. | 7 | Nadruk op scannen en hoofdgedachte achterhalen |
| | **Deel 2** | Een matching-taak met een tekst (artikel, rapport, enz.) met weggelaten zinnen; ongeveer 450-550 woorden. | 5 | Tekststructuur begrijpen |
| | **Deel 3** | A4-optie meerkeuzeopdracht met een tekst; ongeveer 450-550 woorden. | 6 | Hoofdgedachte en specifieke informatie achterhalen |
| | **Deel 4** | A4-optie meerkeuzeopdracht met een informatieve tekst met ontbrekende woorden; tekst bevat gaten; ongeveer 200-300 woorden. | 15 | Woordenschat en structuur |
| | **Deel 5** | Proeflezen. Taak waarbij aanvullende onnodige woorden geïdentificeerd moeten worden in een korte tekst; ongeveer 150-200 woorden. | 12 | Zinsstructuur begrijpen en fouten detecteren |
| **SCHRIJVEN**<br>45 minuten | **Deel 1** | Een bericht, memo of e-mail.<br>Kandidaten worden gevraagd om een interne communicatie te produceren, enkel gebaseerd op de INSTRUCTIES (plus lay-out van de output tekstsoort); 40-50 woorden. | Een verplichte taak | Instructies geven, een ontwikkeling uitleggen, opmerkingen vragen, informatie vragen, akkoord gaan met vragen |
| | **Deel 2** | Zakelijke correspondentie, kort rapport of voorstel<br>Kandidaten worden gevraagd om een stuk zakelijke correspondentie, kort rapport of voorstel, gebaseerd op de instructies en inputtekst(en) te schrijven; 120-140 woorden. | Een verplichte taak | Corresponderen: bv. uitleggen, verontschuldigen, verzekeren, klagen<br>Rapporteren: bv. beschrijven, samenvatten<br>Voorstellen: bv. beschrijven, samenvatten, aanraden, overtuigen |
| **LUISTEREN**<br>40 minuten | **Deel 1** | Een taak met een tekst met gaten waarbij drie korte monologen of dialogen van ongeveer 1 minuut worden gebruikt. Elk fragment wordt twee keer beluisterd. | 12 | Luisteren om notities te nemen |
| | **Deel 2** | Een multiple matching-taak bestaande uit twee delen met telkens vijf korte monologen. | 10 | Luisteren om een thema, context, functie,… te identificeren |
| | **Deel 3** | Een meerkeuzeopdracht met een monoloog, interview of discussie die ongeveer 4 minuten duurt en twee keer wordt beluisterd. | 8 | De hoofdpunten volgen en specifieke informatie uit de tekst halen |
| **SPREKEN**<br>14 minuten | **Deel 1** | Een conversatie tussen de gesprekspartner en de kandidaat (gesproken VRAGEN). | Meerdere | Persoonlijke informatie geven, spreken over hudige omstandigheden, ervaringen uit het verleden en plannen voor de toekomst, mening geven, vertellen wat er zou kunnen gebeuren… |
| | **Deel 2** | Een 'mini-presentatie' door de kandidaat. De kandidaten krijgen de keuze uit drie zakelijke onderwerpen en hebben 1 minuut om een mondelinge prestatie voor te bereiden die ten minste 1 minuut zal duren. | Een presentatie per kandidaat | Een langer stuk mondelinge productie. Informatie geven, mening geven en verantwoorden |
| | **Deel 3** | Een taak waarbij moet worden samengewerkt. Aan de kandidaten wordt een zakelijke discussie aangeboden die de gesprekspartner uitbreidt met responsstimuli over gerelateerde onderwerpen. | Meerdere | Het gesprek beginnen, antwoorden, onderhandelen, samenwerken, informatie uitwisselen, mening geven en verantwoorden, akkoord gaan en/of niet akkoord gaan, suggereren, vertellen wat er zou kunnen gebeuren, vergelijken en contrasteren, beslissingen nemen |

4 Noot van de vertaalster: 'Engels voorbeeld' werd niet toegevoegd aan de vertaalde titel omdat deze volledige handleiding een vertaling is naar het Nederlands van Engelse bronteksten.

# Voorbeeld leesvaardigheid

| Algemene beschrijving | |
|---|---|
| **FORMAT VAN HET TOETSFORMULIER** | Het toetsformulier bevat een reeks zakelijke teksten en bijbehorende taken. Een tekst kan uit verschillende korte delen bestaan. |
| **TIMING** | Een uur. |
| **AANTAL DELEN** | Er zijn vijf delen.<br>Delen 1 tot 3 toetsen het leesbegrip van de kandidaten.<br>Delen 4 en 5 toetsen het begrip van het geschreven Engels op woord-, woordgroep-, zins- en alineaniveau. |
| **AANTAL VRAGEN** | 45 |
| **SOORTEN TAKEN** | Matching.<br>4-opties meerkeuze.<br>4- opties meerkeuze cloze.<br>Proeflezen. |
| **TEKSTSOORTEN** | Informatieve teksten, artikels en rapporten. |
| **LENGTE VAN DE TEKSTEN** | Ongeveer 150-550 woorden per tekst. |
| **FORMAT VOOR DE ANTWOORDEN** | Kandidaten duiden hun antwoord aan door een vakje zwart te kleuren of door een woord op een antwoordformulier te schrijven dat door een computer kan worden gelezen. |
| **SCORES** | Aan elke vraag wordt een punt toegekend. |

# Bijlage IV – Advies voor itemschrijvers

## Advies over hoe teksten te kiezen

De definitie van 'tekst' is in deze handleiding dezelfde als de omschrijving in hoofdstuk 4.6 van het ERK. Het refereert aan om het even welk stuk tekst, zowel geschreven als gesproken.

Itemschrijvers zouden advies moeten krijgen over hoe ze teksten kunnen selecteren. Dit advies zou het volgende moeten bevatten:

- de beste bronnen voor teksten (bv. artikels uit kwaliteitskranten, brochures);
- bronnen die minder geschikte teksten bevatten (bv. gespecialiseerde materialen);
- een algemene waarschuwing om bias te vermijden (bv. wat betreft cultuur, gender, leeftijd, enz.);
- een lijst met redenen waarom teksten in het verleden werden verworpen.

Deze kunnen het volgende bevatten:

- veel culturele en lokale voorkennis (tenzij dit specifiek getoetst wordt);
- onderwerpen die als ongepast kunnen worden beschouwd door de doelgroep, zoals bv. oorlog, dood, politiek en religieuze overtuigingen of andere onderwerpen die sommige kandidaten kunnen schofferen of afleiden;
- onderwerpen waar de kandidaten geen ervaring mee hebben op de leeftijd die ze hebben
- woordenschat of concepten die te moeilijk of te makkelijk zijn;
- technische of stilistische fouten of eigenaardige kenmerken;
- slechte vormgeving van de originele tekst.

Het is ook mogelijk om een lijst met onderwerpen te geven die zo regelmatig aan bod kwamen in eerdere versies van de toets dat ze beter vermeden worden.

In de zoektocht naar een geschikte tekst, kunnen de hoofdstukken 4 en 7 van het ERK een grote hulp zijn omdat ze de teksten helpen situeren in de ruimere context van wat de Raad van Europa taalleren noemt. De media die opgesomd worden in hoofdstuk 4.6.2 (stem, telefoon, radio, enz.) samen met de gesproken en geschreven tekstsoorten die in hoofdstuk 4.6.2 zijn opgenomen, vormen handige checklists en mogelijkheden om soorten items te diversifiëren.

## Advies over de presentatie

Itemschrijvers kunnen richtlijnen krijgen over:

- getypte teksten die al dan niet dubbele interlinie moeten bevatten;
- welke informatie in de hoofding van elke pagina zou moeten staan;
- kopieën of originele teksten die al dan niet moeten worden aangeleverd;
- welke details van de brontekst moeten worden gegeven (bv. publicatiedatum).

## Gedetailleerd advies voor elke taak

Dit kan geïllustreerd worden met een fictief voorbeeld. Dit advies werd bv. gegeven aan een itemschrijver voor een aangepaste cloze-toets die eerder focuste op structuur dan op lexicon:

- Er wordt gevraagd naar een authentieke tekst van ongeveer 200 woorden. De tekst moet een korte titel hebben. De nadruk ligt op enkelvoudige structuurwoorden. Er mogen niet te veel ongekende woorden in voorkomen.
- Er moeten minimaal 16 items zijn, en als het kan liefst meer, zodat er een selectie kan gemaakt worden na de pretest. De eerste items zullen als een voorbeeld gebruikt worden en kunnen genummerd worden met '0' (nul). Items moeten voorzetsels, voornaamwoorden, bepalingen, hulpwerkwoorden, … toetsen. Ze moeten evenredig verspreid worden over de tekst en er moet voor gezorgd worden dat een fout beantwoord item niet automatisch leidt tot een volgend fout antwoord (onafhankelijkheid van de items).
- Meestal is het geen goed idee om het eerste woord in de tekst weg te laten of om een samengestelde vorm weg te laten omdat de kandidaten dan kunnen twijfelen of het om één of meerdere woorden gaat. Een gat dat niet leidt tot een ongrammaticale zin (bv. het woord 'al' weglaten in de volgende zin: 'Ons werd gezegd dat al de treinen vertraging hadden.'), moet worden vermeden, alsook items die focussen op ongewone of eigenaardige kenmerken.

Ook de standaardINSTRUCTIES die bij deze taak horen, kunnen de itemschrijver verder helpen.

Ervaren schrijvers van op een tekst gebaseerde items, verzamelen vaak teksten uit de aangeraden bronnen. Als hen dan gevraagd wordt om items te schrijven, selecteren en bewerken ze de meest beloftevolle teksten uit hun verzameling. Voor het schrijven van bepaalde soorten items (bv. items die focussen op grammatica of woordenschat) is het nuttig voor de itemsschrijvers om een woordenboek of thesaurus bij de hand te hebben. Schrijven ze luistermateriaal, dan is het zinvol om het fragment te beluisteren zodat de items direct gebaseerd kunnen worden op de gesproken tekst en niet op de uitgeschreven tekst.

Veel itemschrijvers vinden het zinvol om materiaal uit te testen door bv. collega's of kennissen die niet bij taaltoetsing betrokken zijn, te vragen om een taak uit te voeren. Dit helpt om fouten te ontdekken zoals typfouten, onduidelijke instructies, foute correctiesleutels en items die heel moeilijk te beantwoorden zijn of waarop meerdere antwoorden mogelijk zijn.

De TOETSSPECIFICATIES zouden ook een soort van checklist moeten bevatten die de itemschrijver kan gebruiken om de tekst, de items en het geheel van de taak na te kijken, voordat die worden ingediend. De checklist die de eerder beschreven cloze-toets begeleidde, wordt hieronder getoond en dient als voorbeeld. Als de tekst, items en taak gepast zijn, dan zou het mogelijk moeten zijn op elke vraag affirmatief te antwoorden.

| **Tekst:** |
|---|
| Is het onderwerp van de tekst toegankelijk/cultureel gezien aanvaardbaar/enz.? |
| Is de tekst vrij van enige mogelijk misplaatste inhoud? |
| Is de tekst op het gepaste moeilijkheidsniveau? |
| Is de tekst geschikt voor een taak die focust op structuur? |
| Is de tekst lang genoeg om een minimum aan 16 items te genereren? |
| Bevat de tekst een gepaste titel? |
| **Items:** |
| Is het gevraagde aantal items gegenereerd? |
| Zijn de items goed verdeeld over de tekst? |
| Is er een goed bereik van de taal waarop gefocust wordt? |
| Is er nagekeken of alle items wel focussen op structuur? |
| Is het zeker dat er geen afhankelijke items zijn? |
| Is er minstens één extra item toegevoegd? |
| Werden items die focussen op eigenaardige kenmerken vermeden? |

| **Instructies en sleutel:** |
| --- |
| Werden de instructies nagekeken? |
| Is er een voorbeeld voorzien? |
| Is er op een apart blad een duidelijke correctiesleutel ingediend? |

Voor er materiaal wordt ingediend, moeten itemschrijvers nagaan of ze van alles een kopie hebben gemaakt. Als de originele teksten van kranten en tijdschriften ingediend worden bij de toetsinstelling, houdt de itemschrijver best gedetailleerde informatie over de bronnen bij op de fotokopies.

# Bijlage V – Case study – een A2-taak reviseren

Deze bijlage toont hoe er veranderingen werden aangebracht aan een taak gedurende het revisieproces en de redenen waarom die veranderingen werden doorgevoerd. Na elke nieuwe versie volgen de opmerkingen. De delen van de tekst die werden besproken, staan in het rood.

## Versie 1 – ingediend door de itemschrijver voor revisie (eerste vergadering)

Vervolledig de conversatie tussen twee vrienden.

Wat zegt Josh tegen zijn vriendin Marta?

Voor vragen **1-5**, duid de juister letter **A-H** aan op uw antwoordenblad.

**VOORBEELD:**

Marta: Hallo Josh. Wat leuk je te zien. Hoe was je vakantie?

Josh: **0** ..... **E**

Antwoord: 0 | A B C D **E** F G H

Marta: Waar ben je dit jaar naartoe geweest?

Josh: **1** .....

Marta: Hoe was het weer?

Josh: **2** .....

Marta: Geweldig! Heb je foto's gemaakt?

Josh: **3** .....

Marta: Heb je in een hotel gelogeerd?

Josh **4** .....

Marta: Zo te horen, heb je een geweldige tijd gehad. Ga je er nog eens heen?

Josh: **5** .....

Marta: Het zou interessanter zijn om dat te doen.

| | |
|---|---|
| **A** | Het was redelijk warm. Ik ben in zee gaan zwemmen. |
| **B** | Nee, dat hebben we niet gedaan. Mijn oom heeft daar vrienden wonen en wij hebben bij hen gelogeerd. |
| **C** | Ik vond ze echt geweldig. |
| **D** | Waarschijnlijk niet. Ik denk volgend jaar naar een andere plek te gaan. |
| **E** | Het was geweldig. Dank je. |
| **F** | Ik heb er echt een paar leuke gemaakt. |
| **G** | We hadden niet genoeg geld. |
| **H** | Ik ben met mijn oom naar IJsland geweest. |

Sleutel: 1H, 2A, 3F, 4B, 5D

## Controle van de versie die voor revisie werd ingediend (eerste vergadering)

Tijdens de (eerste) vergadering over de revisie, werd de schrijver gevraagd om de taak opnieuw in te dienen, met de volgende veranderingen:

- vermijd het voortdurende antwoord/vraag-patroon doorheen de conversatie;
- vermijd de herhaling van bepaalde woorden;
- verbeter de afleider **G** en **C** en de gerelateerde tekst;
- herformuleer de woorden en structuren die niet tot de woordenlijst en grammaticale TOETSSPECIFICATIES horen.

De eerste aanpassing was nodig om ervoor te zorgen dat de taak niet te makkelijk werd en om te focussen op zichzelf staande antwoorden en vragen. In de originele versie werd er met elk opengelaten stuk gekeken wat Josh zijn respons was op een vraag van Marta. Er werd de schrijver gevraagd om variatie te brengen in het interactiepatroon (bv. door sommige opties **A-H** in vragen te veranderen) en om delen te herformuleren (bv. door een aanbod toe te voegen aan optie **F**) zodat er in de dialoog meer samenhang zou ontstaan.

De tweede aanpassing werd doorgevoerd om te vermijden dat dezelfde werkwoordsvorm in zowel de vraag als het antwoord voorkwam en de taak daardoor te makkelijk werd. 'Heb je foto's gemaakt? ' en 'Ik heb er echt een paar leuke gemaakt.'; 'Waar ben je dit jaar naartoe geweest?' en 'Ik ben met mijn oom naar IJsland geweest.' Er werd aan de schrijver ook gevraagd om de woordenschat te variëren. 'Geweldig' kwam bijvoorbeeld vier keer voor.

De derde aanpassing werd gedaan omdat de afleiders **C** en **G** mogelijke antwoorden waren voor meerdere items. Er werd de schrijver gevraagd om **C** en **G** en de gerelateerde tekst te herformuleren zodat zij niet bij item 3 zouden passen en om ervoor te zorgen dat **G** niet te aanlokkelijk was voor item 4.

De vierde aanpassing had te maken met de moeilijkheidsgraad van de inhoud van de taak. Aan de schrijver werd bijvoorbeeld gevraagd om 'waarschijnlijk' te vervangen omdat het niet in de woordenlijst stond en 'Het zou…zijn…' te herformuleren omdat het niet in de lijst met functies voor deze toets stond.

# Versie 2 – De herschreven taak die opnieuw werd ingediend door de itemschrijver

Veranderingen aangebracht door de itemschrijver:

**i.** het interactiepatroon is gevarieerder, 'waarschijnlijk' en 'zou… zijn…' zijn verwijderd.

**ii.** Optie C is veranderd zodat het niet meer werkt voor item 3.

**iii.** De tekst voor en na de opening voor item 4 werd verbeterd om G uit te sluiten voor items 3 en 4.

**iv.** Het werkwoord 'gemaakt' werd verwijderd uit optie F.

**v.** De itemschrijver argumenteerde dat door 'ben geweest' te vervangen door bijvoorbeeld 'heb…bezocht' de dialoog onnatuurlijk zou worden, dus de twee vormen van 'zijn' werden in langere segmenten van de tekst verwerkt.

Vervolledig de conversatie tussen twee vrienden.

Wat zegt Josh tegen zijn vriendin Marta?

Voor vragen **1-5**, duid de juister letter **A-H** aan op uw antwoordenblad.

**VOORBEELD:**

Marta: Hallo Josh. Wat leuk je te zien. Hoe was je vakantie?

Josh: **0** ..... **E**

Antwoord:

| 0 | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| | ☐ | ☐ | ☐ | ☐ | ■ | ☐ | ☐ | ☐ |

Marta: Waar ben je dit jaar naartoe geweest? Terug naar je oom?

Josh: **1** .....

Marta: Nee, het is daar veel te koud voor mij.

Josh: **2** .....

Marta: Geweldig! Heb je foto's gemaakt?

Josh: **3** .....

Marta: Ja, graag! Heb je in een hotel gelogeerd?

Josh **4** .....

Marta: Jullie hadden geluk!

Josh: **5** .....

Marta: Dat wist ik niet. Je moet me er binnenkort alles over vertellen.

| | |
|---|---|
| **A** | Nee, in de zomer is het er redelijk warm. Je kan er zelfs in zee gaan zwemmen. |
| **B** | Mijn oom heeft daar vrienden wonen en wij hebben bij hen gelogeerd. |
| **C** | Nee, maar heb jij van je vakantie genoten? |
| **D** | Dat hadden we zeker. De hotels zijn daar echt duur. |
| **E** | Het was leuk. Dank je. |
| **F** | Super veel. Als je wilt, breng ik ze mee om ze je te laten zien. |
| **G** | We hadden niet genoeg geld. |
| **H** | We hebben iets anders gedaan: wij zijn naar IJsland gegaan. Ben je daar al geweest? |

Sleutel: 1H, 2A, 3F, 4B, 5D

# Versie 2 – De herschreven taak die opnieuw werd ingediend door de itemschrijver na de revisiebespreking (tweede vergadering)

Vervolledig de conversatie tussen twee vrienden.

Wat zegt Josh tegen zijn vriendin Marta?

Voor vragen **1-5**, duid de juister letter **A-H** aan op uw antwoordenblad.

**VOORBEELD:**

Marta: Hallo Josh. Wat leuk je te zien. Hoe was je vakantie?

Josh: **0** ..... **E**

Antwoord: 0 | A B C D **E** F G H

Marta: Waar ben je dit jaar naartoe geweest? Terug naar je oom?

Josh: **1** .....

Marta: Nee, het is daar veel te koud voor mij.

Josh: **2** .....

Marta: Geweldig! Heb je foto's gemaakt?

Josh: **3** .....

Marta: Ja, graag! Heb je in een hotel gelogeerd?

Josh **4** .....

Marta: Jullie hadden geluk!

Josh: **5** .....

Marta: Dat wist ik niet. Je moet me er binnenkort alles over vertellen.

| | |
|---|---|
| **A** | Nee, in de zomer is het er redelijk warm. Je kan er zelfs in zee gaan zwemmen. |
| **B** | Mijn oom heeft daar vrienden wonen en wij hebben bij hen gelogeerd. |
| **C** | Nee, maar heb jij van je vakantie genoten? |
| **D** | Dat hadden we zeker. De hotels zijn daar echt duur. |
| **E** | Het was leuk. Dank je. |
| **F** | Super veel. Als je wilt, breng ik ze mee om ze je te laten zien. |
| **G** | We hadden niet genoeg geld. |
| **H** | We hebben iets anders gedaan: wij zijn naar IJsland gegaan. Ben je daar al geweest? |

Sleutel: 1H, 2A, 3F, 4B, 5D

## Controle van de versie die opnieuw werd ingediend voor revisie (tweede vergadering)

Tijdens de (tweede) vergadering over de revisie van deze taak werden de volgende veranderingen doorgevoerd:

- 'zijn vriendin' werd uit de tweede lijn van de INSTRUCTIES verwijderd.
- Marta's tweede beurt (tussen opening 1 en 2) werd veranderd (in 'Nee. Is het daar niet erg koud?').
- Optie **A** werd veranderd in 'Niet echt. In de zomer kan je er zelfs in zee zwemmen'.
- Optie **B** werd veranderd in 'We hebben daar vrienden wonen en we hebben bij hen thuis geslapen'.

De eerste aanpassing was van stilistische aard om te vermijden dat het woord 'vriend' werd herhaald, dat voorkwam in de eerste lijn van de instructies, en om de instructies te standaardiseren.

Er waren twee redenen om Marta's tweede beurt te veranderen. Ten eerste moest optie **B** worden uitgesloten voor item 2 (de sleutel is **A**). Ten tweede moest het duidelijker zijn wat er inhoudelijk moest worden ingevuld. In de aanvankelijke versie was het mogelijk dat de focus van de conversatie veranderde na Marta's beurt, maar door dit in een vraag te veranderen, werd er meer aandacht gegeven aan het antwoord van Josh.

Optie **A** werd dan veranderd omdat Marta's tweede beurt een vraag was geworden. 'Niet echt' werd een respons op Marta's vraag en de informatie over de zomer en zwemmen in zee werd behouden, maar een beetje verbeterd.

Er waren ook twee redenen waarom optie **B** werd veranderd, die de sleutel is voor item 4. Eerst werd de referentie aan de oom van Josh verwijderd omdat dit aanzette om de optie aan het begin van de conversatie te plaatsen en niet zozeer op de vierde open lijn in te vullen. Het verwijswoord 'hen' in **B** kon ook verwarrend zijn. Daarom werd er beslist om een onderscheid te maken tussen het huis dat toebehoorde aan de familie van zijn oom en het huis dat van de vrienden van zijn oom was. Als de kandidaten dit onderscheid niet hadden opgemerkt, dan hadden ze optie **B** niet overwogen voor item 4. Optie B werd daarom veranderd in 'We hebben daar vrienden wonen'. De tweede reden waarom optie B werd aangepast, was om te vermijden dat er puur op lexicaal niveau zou worden gematcht. 'Gelogeerd' komt voor in Marta's tweede vraag voor item 4 en het komt ook voor in optie **B**. 'We hebben bij hen gelogeerd' werd daarom veranderd in 'We hebben bij hen thuis geslapen'.

# Versie 3 – De versie voor de proefafname – bevat de veranderingen die tijdens de tweede vergadering werden doorgevoerd

Vervolledig de conversatie tussen twee vrienden.

Wat zegt Josh tegen Marta?

Voor vragen **1-5**, duid de juister letter **A-H** aan op uw antwoordenblad.

**VOORBEELD:**

Marta: Hallo Josh. Wat leuk je te zien. Hoe was je vakantie?

Josh: **0** ..... **E**

Antwoord:

| 0 | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| | ☐ | ☐ | ☐ | ☐ | ■ | ☐ | ☐ | ☐ |

Marta: Waar ben je dit jaar naartoe geweest? Terug naar je oom?

Josh: **1** .....

Marta: Nee. Is het daar niet erg koud?

Josh: **2** .....

Marta: Geweldig! Heb je foto's gemaakt?

Josh: **3** .....

Marta: Ja, graag! Heb je in een hotel gelogeerd?

Josh **4** .....

Marta: Jullie hadden geluk!

Josh: **5** .....

Marta: Dat wist ik niet. Je moet me er binnenkort alles over vertellen.

| | |
|---|---|
| **A** | Niet echt. In de zomer kan er je zelfs in zee zwemmen. |
| **B** | We hebben daar vrienden wonen en we hebben bij hen thuis geslapen. |
| **C** | Nee, maar heb jij van je vakantie genoten? |
| **D** | Dat hadden we zeker. De hotels zijn daar echt duur. |
| **E** | Het was leuk. Dank je. |
| **F** | Super veel. Als je wilt, breng ik ze mee om ze je te laten zien. |
| **G** | We hadden niet genoeg geld. |
| **H** | We hebben iets anders gedaan: wij zijn naar IJsland gegaan. Ben je daar al geweest? |

Sleutel: 1H, 2A, 3F, 4B, 5D

# Controle van de gepreteste versie (derde vergadering)

Er werd tijdens deze vergadering beslist dat de taak niet verder moest worden aangepast. De beschrijvende statistieken toonden aan dat de moeilijkheidsgraad van de taak correct was (zie Bijlage VII voor een beschrijving van hoe deze statistieken moeten begrepen worden). De beoogde gemiddelde moeilijkheidsgraad voor de toets Cambridge English: Key (KET)[5] is -2.09 en deze taak had een gemiddelde moeilijkheidsgraad van -2.31. Items 1-5 vielen binnen het aanvaardbare BEREIK van moeilijkheid, dat tussen -3.19 en -0.99 ligt.

| | Moeilijkheid van het item (LOGITS) |
|---|---|
| **1** | -2.72 |
| **2** | -2.90 |
| **3** | -2.86 |
| **4** | -1.92 |
| **5** | -1.13 |
| **GEMIDDELDE** | -2.31 |

Mogelijk dubbele antwoorden uitsluiten was ook een aandachtspunt tijdens deze vergadering. De uitsplitsing van de antwoorden van de kandidaten wordt getoond in het rapport van de statistische analyse hieronder. Voor item 2 bijvoorbeeld, koos 2,20% van de laag scorende groep voor **F** en voor item 4 koos 4,50% van de laag scorende groep voor **D**. Deze opties werden opnieuw gecontroleerd om te zien of ze antwoorden op 2 en 4 zouden kunnen zijn. Er werd beslist dat dit niet mogelijk was. Optie **F** kon bijvoorbeeld geen antwoord zijn voor item 2 omdat 'Als je wilt, breng ik ze mee om ze je te laten zien' aan niks daaruit refereert. Optie **D** is ook uitgesloten als sleutel voor item 4 door 'Dat hadden we zeker'.

| Item Statistics | | | | | Alternative Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Seq No. | Scale -Item | Prop. Correct | Disc. Index | Point Biser. | Alt. | Prop. Total | Endorsing Low | Endorsing High | Point Biser. | Key |
| 1 | 1-13 | .73 | .41 | .40 | A | .07 | .15 | .00 | -.24 | |
| | | | | | B | .07 | .15 | .00 | -.27 | |
| | | | | | C | .09 | .11 | .02 | -.13 | |
| | | | | | D | .01 | .00 | .00 | -.04 | |
| | | | | | E | .01 | .00 | .00 | -.06 | |
| | | | | | F | .01 | .00 | .00 | .05 | |
| | | | | | G | .02 | .02 | .00 | -.04 | |
| | | | | | H | .73 | .57 | .98 | .40 | * |
| | | | | | Other | .00 | .00 | .00 | | |
| 2 | 1-14 | .76 | .41 | .36 | A | .76 | .57 | .98 | .36 | * |
| | | | | | B | .04 | .07 | .00 | -.08 | |
| | | | | | C | .05 | .07 | .00 | -.11 | |
| | | | | | D | .03 | .04 | .00 | -.06 | |
| | | | | | E | .03 | .04 | .00 | -.15 | |
| | | | | | F | .07 | .20 | .02 | -.30 | |
| | | | | | G | .00 | .00 | .00 | | |
| | | | | | H | .02 | .02 | .00 | -.04 | |
| | | | | | Other | .00 | .00 | .00 | | |
| 3 | 1-15 | .76 | .41 | .38 | A | .02 | .00 | .00 | .00 | |
| | | | | | B | .03 | .00 | .00 | -.02 | |
| | | | | | C | .07 | .15 | .04 | -.15 | |
| | | | | | D | .03 | .09 | .02 | -.16 | |
| | | | | | E | .04 | .11 | .00 | -.21 | |
| | | | | | F | .76 | .50 | .91 | .38 | * |
| | | | | | G | .04 | .11 | .02 | -.18 | |
| | | | | | H | .01 | .02 | .00 | -.15 | |
| | | | | | Other | .01 | .00 | .00 | -.08 | |
| 4 | 1-16 | .58 | .56 | .45 | A | .01 | .00 | .00 | .01 | |
| | | | | | B | .58 | .28 | .84 | .45 | * |
| | | | | | C | .02 | .00 | .04 | .12 | |
| | | | | | D | .29 | .50 | .07 | -.37 | |
| | | | | | E | .01 | .02 | .00 | -.04 | |
| | | | | | F | .00 | .00 | .00 | | |
| | | | | | G | .10 | .20 | .04 | -.22 | |
| | | | | | H | .00 | .00 | .00 | | |
| | | | | | Other | .00 | .00 | .00 | | |
| 5 | 1-17 | .41 | .60 | .50 | A | .02 | .02 | .00 | -.01 | |
| | | | | | B | .07 | .09 | .07 | -.06 | |
| | | | | | C | .10 | .07 | .07 | -.05 | |
| | | | | | D | .41 | .13 | .73 | .50 | * |
| | | | | | E | .17 | .35 | .02 | -.33 | |
| | | | | | F | .06 | .09 | .07 | -.04 | |
| | | | | | G | .07 | .11 | .04 | -.06 | |
| | | | | | H | .09 | .15 | .00 | -.22 | |
| | | | | | Other | .00 | .00 | .00 | | |

5 Noot van de vertaalster.

## Versie 4 – De versie voor de live toets
## (zelfde als versie 3 – er werd niks aan veranderd)

# Bijlage VI – Informatie verzamelen uit pretests/proefafnames

Deze bijlage bevat vragen die kunnen worden gesteld na het PRETESTEN of de PROEFAFNAME (zie hoofdstuk 3.4.2).

## Feedback van toezichthouders van de pretest/proefafname – alle onderdelen

Bespreek a.u.b. het volgende:

1. **Inhoud**: het bereik en de soorten vragen/teksten/taken, enz.
2. **Niveau**: moeilijkheidsgraad van de verschillende onderdelen/taken op bv. linguïstisch/cognitief vlak.
3. **Enkel voor pretests van luistervaardigheidstoetsen**: helderheid/snelheid van de opname, accent van de sprekers, enz.
4. **Kandidaten**: welke leeftijd hebben kandidaten die aan de pretest deelnamen ongeveer?
5. **Andere opmerkingen?**

## Feedback van kandidaten van de pretest/proefafname – leestoets

1. **Had u genoeg tijd om de taak uit te voeren? (Indien dit niet het geval was, hoeveel meer tijd zou u nodig gehad hebben?)**
2. **Waren er woorden in de taak die u niet begreep?**
   (Noteer a.u.b. de specifieke woorden/uitdrukkingen die een probleem vormden.)
3. **Kon u de ideeën en de argumentatie van de schrijver volgen?**
   GEMAKKELIJK/MET ENIGE MOEITE/MET VEEL MOEITE
4. **Hoe goed was u vertrouwd met het onderwerp van de passage?**
   HEEL VERTROUWD/REDELIJK VERTROUWD/NIET ECHT VERTROUWD/HELEMAAL NIET VERTROUWD
5. **Wanneer (als dit het geval zou zijn) bent u van plan de live toets af te leggen?**
6. **Heeft u andere opmerkingen?**

## Feedback van beoordelaars van de pretest/proefafname – schrijftoets

Wat betreft de **INPUT** van de taak: focus

1. Werd de taak in het algemeen begrepen?
2. Werd de rol van de schrijver duidelijk herkend?
3. Werd de beoogde lezer duidelijk geïdentificeerd?
4. Is er sprake van culturele bias? Bevoordeelt de taak bepaalde kandidaten op basis van achtergrond of leeftijd?
5. Moeten de vragen anders geformuleerd worden? Als dit zo is, kunt u dan zeggen hoe dit zou kunnen?

Wat betreft de **INPUT** van de taak: taal

6. Waren de vragen begrijpelijk voor de kandidaten met een B2-niveau?
7. Was er enige verwarring over of een verkeerde interpretatie van de gebruikte woorden?
8. Twijfelden de kandidaten over het geschikte REGISTER om te gebruiken?
9. Moeten de vragen anders geformuleerd worden? Als dit zo is, kunt u dan zeggen hoe dit zou kunnen?

Wat betreft de **OUTPUT** van de taak: inhoud

10. Werd het soort taak juist geïnterpreteerd?
11. Werden er inhoud onderdelen verkeerd begrepen/weggelaten? Geef a.u.b. details.
12. Was het aantal woorden geschikt voor de taak?

Wat betreft de **OUTPUT** van de taak: bereik/toon

13. Werden er talige elementen uit de vraag overgenomen? Specifieer a.u.b.
14. Welk register (formeel, informeel,…) hanteerden de kandidaten?

Wat betreft de **OUTPUT** van de taak: niveau

15. Liet de vraag kandidaten met een C1-niveau voldoende toe om hun vaardigheden te tonen?

**Correctieschema**:

16. Geef a.u.b. suggesties voor de verbetering van het correctieschema.

**Algemene indruk:**

17. Geef a.u.b. uw algemene indruk van de vraag.

# Bijlage VII – Statistische informatie gebruiken in de toetscyclus

Toetsdata verzamelen en analyseren vraagt de nodige planning en middelen, maar de kwaliteit van de toets en de interpreteerbaarheid van de uitkomsten worden er wel aanzienlijk door verhoogd. Er zou ten minste informatie moeten worden verzameld over de kandidaten die de toets hebben afgelegd en de scores die ze hebben behaald. Deze informatie kan eenvoudig worden samengevat door simpele beschrijvende statistieken (zie bijvoorbeeld Carr, 2008).

Meer precieze data over de prestaties van de kandidaten kunnen aantonen hoe goed items gewerkt hebben en aan welke aspecten er bij de revisie aandacht moet worden besteed. Er bestaan gebruiksvriendelijke softwarepakketten waarmee de analyses kunnen worden uitgevoerd zoals hieronder beschreven. Die pakketten kunnen gebruikt worden wanneer er met een kleine groep kandidaten wordt gewerkt. (bv. 50).

Deze aanvullende gegevens kunnen worden verzameld:

- taakniveau: de score die de kandidaat voor elke taak heeft behaald, en niet alleen de totale score.
- RESPONSEN: de antwoorden die de kandidaat op elk item uit de toets gaven.
- demografische informatie: informatie over de kandidaat zoals leeftijd, gender, eerste taal, enz.

## De data

De meeste analysesoftware vraagt gegevens die er uit zien als in Figuur 18. U kan om het even welke tekstverwerker gebruiken om de data in te voeren, maar:

- gebruik een font met een vaste breedte zoals Courier;
- gebruik geen tabs;
- bewaar het document als niet-opgemaakte tekst (.txt).


**Figuur 18** ***Een typisch document met responsdata***

In Figuur 18:

- bevat elke rij de responsen van slechts één kandidaat;
- bevat de eerste kolom het identificatienummer van de kandidaat (dit kan ook demografische informatie zijn);

- bevat elke kolom de antwoorden op slechts één toetsitem.

Dit is een voorbeeld voor een meerkeuzetoets waar de opties (a-h) worden genoteerd die door elke kandidaat werden geselecteerd.

De analysesoftware zal wat aanvullende informatie nodig hebben zoals welke optie de correcte is voor elk item.

## Klassieke itemanalyse

Klassieke ITEMANALYSE wordt gebruikt:

- om pretestdata te analyseren en informatie te geven voor de selectie en revisie van taken die in de live toets komen;
- om de responsdata uit de live toets te analyseren.

Het geeft een waaier aan statistische gegevens over de werking van de items en de toets als geheel. In het bijzonder:

Beschrijvende statistieken die iets zeggen over de werking van elk item:

- hoe makkelijke een item was voor een groep kandidaten;
- hoe goed een item discrimineert tussen sterke en zwakke kandidaten;
- hoe goed de sleutel en elke afleider heeft gewerkt.

Samenvattende statistieken voor de gehele toets, of voor elk onderdeel, onder andere:

- het aantal kandidaten;
- het gemiddelde en de STANDAARDDEVIATIE van de scores;
- een schatting van de betrouwbaarheid.

Hieronder volgen enkele richtlijnen over wat aanvaardbare waarden zijn voor sommige van deze statistieken. Ze kunnen niet als absolute regels worden beschouwd, want in de praktijk zijn de waarden die normaal worden geobserveerd afhankelijk van de context. De beschrijvende statistieken leveren vaker aanvaardbare waarden op als:

- u een groot aantal items hebt in een toets;
- u meer kandidaten hebt;
- het bereik van de vaardigheden van de groep die de toets aflegt groter is

en omgekeerd zien ze er minder goed uit als u minder items of kandidaten hebt of het bereik van de vaardigheden beperkt is.

Figuur 19 is een voorbeeld van itemstatistieken, gebaseerd op de output van het MicroCAT itemanalysepakket (zie Tools voor statistische analyse hieronder). Het toont de analyse voor drie items.

| | Item Statistics | | | Alternative Statistics | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Seq. No. | Scale -Item | Prop. Correct | Disc. Index | Point Biser. | Alt. | Prop. Total | Endorsing Low | Endorsing High | Point Biser. | Key |
| 1 | 1-1 | .38 | .52 | .48 | A | .00 | .00 | .00 | | |
| | | | | | B | .38 | .13 | .66 | .48 | * |
| | | | | | C | .12 | .11 | .12 | -.01 | |
| | | | | | D | .49 | .74 | .23 | -.44 | |
| | | | | | Other | .01 | .00 | .00 | -.11 | |
| 2 | 1-2 | .71 | .42 | .42 | A | .07 | .11 | .01 | -.16 | |
| | | | | | B | .11 | .18 | .04 | -.22 | |
| | | | | | C | .10 | .16 | .00 | -.22 | |
| | | | | | D | .71 | .53 | .95 | .42 | * |
| | | | | | Other | 01 | .00 | .00 | -.13 | |
| 3 | 1-3 | .93 | .19 | .39 | A | .93 | .81 | .00 | .39 | * |
| | | | | | B | .07 | .18 | .00 | -.39 | |
| | | | | | Other | .01 | .00 | .00 | -.03 | |

**Figuur 19** ***Voorbeeld van itemstatistieken (MicroCAT itemanalysepakket)***

## Proportie correct

PROPORTIE CORRECT is het aantal juiste responsen (*Prop. Correct* in Figuur 19). Het toont hoe makkelijk een item was voor de groep van kandidaten. De waarde ligt tussen 0 en 1, waarbij een hoger cijfer staat voor een makkelijker item. Figuur 19 toont dat item 1 het moeilijkst is en item 3 het makkelijkst.

De proportie correct is de eerste beschrijvende statistiek om te bekijken, want als die te hoog of te laag is (bv. buiten het bereik 0.25-0.80%) dan betekent dit dat andere statistieken niet goed werden geschat. Met andere woorden: we beschikken niet over goede informatie van deze groep kandidaten. Als ze representatief zijn voor de groep die de live toets zal afleggen, dan kunnen we besluiten dat het item gewoonweg te gemakkelijk of te moeilijk is. Als we niet zeker zijn van het niveau van de kandidaten, dan zou het kunnen dat het item in orde is, maar dat de groep niet het juiste niveau heeft. Een praktisch besluit is dat we altijd zouden moeten proberen om pretests te organiseren met groepen die ruwweg hetzelfde taalniveau hebben als de reëele toetspopulatie.

## Discriminatie

Goede items zouden zwakkere en sterkere kandidaten van elkaar moeten onderscheiden. Klassieke itemanalyse kan hier twee indicaties van geven: de DISCRIMINATIE-index en de punt biseriële CORRELATIE-coëfficiënt (*Disc. Index* en *Point Biser.* in Figuur 19).

De discriminatie-index is een eenvoudige statistiek: het is het verschil tussen de proportie juiste antwoorden die gegeven werden door de hoogst scorende kandidaten en de proportie juiste antwoorden gegeven door de laagst scorende kandidaten (meestal het bovenste en onderste derde van de kandidaten). De output in Figuur 19 toont dit in de kolommen *Low* (laag) en *High* (hoog). Voor item 1 is het verschil tussen de hoge en lage groep (0.66-0.13). Dit is de waarde van de discriminatie-index (na afronding).

Een item dat goed discrimineert, heeft een discriminatiewaarde die +1 benadert en toont dat de sterkste kandidaten dit item doorgaans juist hebben terwijl de zwakste kandidaten het doorgaans fout hebben.

Als de proportie correct heel hoog of heel laag is, zullen zowel de lage als hoge groepen goed (of slecht) scoren. De discriminatie zal door deze index dus onderschat worden. Item 3 illustreert dit probleem: 1.00-0.81 = 0.19, een lage waarde.

De punt biseriële correlatiecoefficiënt houdt een moeilijkere berekening in dan de discriminatie-index en wordt minder beïnvloed door een hoge of lage proportie correct. Het is een correlatie tussen de scores van kandidaten op een item (1 of 0) en op de toets in zijn geheel.

Over het algemeen worden items met een punt biseriële correlatiecoëfficiënt die groter is dan 0.30 als acceptabel beschouwd. Een negatieve punt biseriële correlatiecoëfficiënt betekent dat sterke kandidaten waarschijnlijk meer kans maken om het item fout te hebben. In dit geval moet er worden gecontroleerd of één van de afleiders eigenlijk een goed antwoord is of de correctiesleutel fout is.

## Afleideranalyse

Afleiders zijn de foute opties in een meerkeuze-item. We verwachten dat zwakke kandidaten de afleider zullen selecteren, terwijl sterke kandidaten de correctie optie zullen aanduiden (*Key* in Figuur 19, aangeduid met '*').

Een afleideranalyse toont het aantal kandidaten dat voor elke afleider heeft gekozen (*Prop. Total* in Figuur 19). Item 1 in Figuur 19 heeft een redelijk lage proportie correct (optie B): 0.38. De afleider D werd vaker geselecteerd: 0.49[6]. Afleider A werd door geen enkele kandidaat gekozen, dus het is duidelijk geen goede afleider. Maar in het algemeen werkt het item goed, en discrimineert het goed, dus het is niet nodig om het te veranderen. In de praktijk is het moeilijk om drie afleiders te vinden die allemaal goed werken.

De analyse in Figuur 19 toont ook de proportie van de hoogst en de laagst scorende kandidaten die elke optie kozen en de punt biseriële correlatiecoefficiënt voor elke optie. Een goed item zal een positieve punt biseriële correlatiecoëfficiënt hebben voor de correcte optie en een negatieve voor elke afleider.

## Betrouwbaarheid van de scores

Er zijn meerder manieren om de betrouwbaarheid te schatten en verschillende formules die hierbij kunnen worden gebruikt. Elke methode gaat uit van verschillende aannames. Met de *split half*-methode wordt de toets in twee equivalente delen verdeeld en worden de scores van de kandidaten op de verschillende delen met elkaar vergeleken. Bij deze methode is het belangrijk dat de twee helften zo equivalent mogelijk zijn: zij omvatten equivalente elementen van het construct, hun moeilijkheidsgraad is equivalent,…

Andere methodes meten de interne consistentie van de toets. Dit werkt goed als de items heel vergelijkbaar zijn in type en inhoud. Maar wanneer items heterogeen zijn, zal de betrouwbaarheid onderschat worden.

Voor klassieke itemanalyse:

**minimaal aantal kandidaten**: 50 tot 80 (Jones, Smith en Talley, 2006, p. 495)

**meer informatie**: Verhelst (2004a,b); Bachman (2004)

## Rasch-analyse

RASCH-ANALYSE is de eenvoudigste en de meest praktische vorm van ITEMRESPONSTHEORIE of IRT. Het zorgt voor een beter begrip van de moeilijkheidsgraad van de items dan de klassieke itemanalyse en heeft meer aanvullende toepassingen, zoals het LINKEN van verschillende versies van de toets.

Met een Rasch-analyse:

- is het precieze verschil in moeilijkheidsgraad tussen twee items duidelijk, want items worden op een INTERVALSCHAAL gezet, uitgedrukt in logits;
- het verschil tussen items en kandidaten, toetsscores of de grens tussen geslaagden en niet-geslaagden kan begrepen worden op dezelfde manier, al deze dingen worden op eenzelfde schaal gemeten;
- de moeilijkheid van een item kan begrepen worden onafhankelijk van de invloed die wordt uitgeoefend door de vaardigheden van de kandidaat (met klassieke itemanalyse zou een groep van vaardige kandidaten een item makkelijk kunnen doen lijken, of een zwakke groep zou het moeilijk kunnen laten lijken).

Deze eigenschappen betekenen dat de Rasch-analyse nuttig is om standaarden te monitoren en te behouden, over sessies heen. Als de Rasch-analyse op deze manier gebruikt wordt, moeten items uit verschillende toetsen wel met elkaar gelinkt worden. Twee toetsen kunnen bijvoorbeeld op deze manieren gelinkt worden:

6 Noot van de vertaalster: In de oorspronkelijke versie wordt foutief naar deze proportie verwezen als de facility (= proportie correct).

- sommige items worden in beide toetsen gebruikt;
- een groep ANKERITEMS wordt in beide toetsen gebruikt;
- sommige of alle items worden gekalibreerd voor ze in een live toets gebruikt worden (KALIBRATIE; zie hoofdstuk 3.4.2, pretest);
- sommige kandidaten maken beide toetsen.

Wanneer de data van beide toetsen zijn geanalyseerd, vormt de link een referentiekader voor alle items, kandidaten, enz. en de items krijgen gekalibreerde moeilijkheidswaarden. Andere toetsen kunnen op dezelfde manier aan dit referentiekader toegevoegd worden.

Standaarden kunnen gemonitord worden door de relatieve positie van belangrijke elementen te vergelijken:

- Hebben de items van alle toetsen hetzelfde moeilijkheidsniveau?
- Zijn de vaardigheden van de kandidaten dezelfde?
- Komen de grenzen tussen geslaagden en niet-geslaagden (gemeten in logits) overeen met dezelfde RUWE SCORE (nu ook gemeten in logits) in alle versies van de toets?

Standaarden kunnen worden behouden als de grenzen tussen geslaagden en niet-geslaagden telkens op dezelfde moeilijkheidswaarden worden gezet.

Het is echter makkelijker om standaarden en de kwaliteit van toetsen te behouden als de toetsen samengesteld zijn uit gekalibreerde items. De algemene moeilijkheidsgraad van een toets kan worden beschreven aan de hand van zijn gemiddelde moeilijkheid en het BEREIK. De moeilijkheid van een toets kan onder controle worden gehouden door een groep items te selecteren die binnen het bedoelde bereik liggen en overeenkomen met het beoogde gemiddelde.

Als u items begint te kalibreren, zullen de moeilijkheidswaarden weinig betekenen. Maar na verloop van tijd kunt u de vaardigheid van de kandidaten in de echte wereld bestuderen om betekenis te geven aan de punten op de vaardigheidsschaal. Daarnaast, of tegelijkertijd, kunt u subjectieve oordelen vellen over de items ('Ik denk dat een threshold B1-leerder 60% kans zou hebben om dit item juist te hebben.'). Op die manier worden de cijfers op de vaardigheidsschaal herkenbaar en betekenisvol.

Voor Rasch-analyse:

**minimaal aantal kandidaten:** 50 tot 80 (Jones, Smith en Talley, 2006, p. 495)

**meer informatie**: Verhelst (2004d); Bond en Fox (2007)

## Statistieken voor correctie en beoordeling

### Methodische correctie

Het is belangrijk dat de correctoren goed presteren. Als dit niet het geval is, moet er actie (zoals een nieuwe training) ondernomen worden (zie hoofdstuk 5.1). Als het om een klein aantal kandidaten gaat, is het misschien mogelijk om na te gaan hoe elke corrector elk item heeft gecorrigeerd. Als het om een grotere groep kandidaten gaat, kan er een steekproef (misschien 10%) van hun werk worden nagekeken. Op basis daarvan kan dan het foutenpercentage bepaald worden. Het foutenpercentage is het aantal fouten dat de corrector heeft gemaakt, gedeeld door het aantal gecorrigeerde items. Als de steekproef representatief is voor al het werk van een corrector, dan is het foutenpercentage waarschijnlijk soortgelijk voor alle correcties door de corrector.

Om ervoor te zorgen dat de steekproef representatief is voor het werk van de corrector, wordt die best willekeurig samengesteld. Om er zeker te kunnen van zijn dat het om een echt willekeurige steekproef gaat, is het belangrijk om stil te staan bij de manier waarop de corrector te werk is gegaan. Een willekeurige steekproef betekent niet dat om het even welke 10% van het werk kan worden gecontroleerd, want dan zou enkel recent werk meegerekend kunnen worden omdat dit toegankelijker is. In dit geval zou het foutenpercentage onderschat kunnen worden voor de hele periode die de correctie behelst: de corrector zal waarschijnlijk beter geworden zijn gedurende het correctieproces.

## Beoordelen

De prestatie van de BEOORDELAARS kan op een heel eenvoudige manier statistisch geëvalueerd worden door het gemiddelde te berekenen van de punten die ze hebben toegekend en de STANDAARDDEVIATIE (een maat voor de spreiding van hun scores, van laag naar hoog). Verschillende beoordelaars kunnen met elkaar vergeleken worden en wie anders blijkt te zijn dan de anderen kan verder geëvalueerd worden. Dit zal werken als het toetsmateriaal willekeurig aan de beoordelaars is toegekend. Als dit niet het geval is, dan kan een beoordelaar gevraagd worden om een oordeel te vellen over kandidaten die normaal beter of slechter zijn dan het gemiddelde. In dit geval zal het gemiddelde hoger of lager liggen dan het gemiddelde van de andere beoordelaars, maar het werk van de beoordelaar kan wel goed zijn.

Als sommige taken door twee beoordelaars worden beoordeeld, dan kan de betrouwbaarheid van deze scores worden geschat. Dit kan bijvoorbeeld door in Excel gebruik te maken van de 'Pearson' correlatiefunctie. De data kunnen als volgt klaargezet worden:

| | beoordelaar 1 | beoordelaar 2 |
|---|---|---|
| kandidaat 1 | 5 | 4 |
| kandidaat 2 | 3 | 4 |
| kandidaat 3 | 4 | 5 |
| … | … | … |

De correlatiecoëfficiënt zal tussen -1 en 1 liggen. In de meeste gevallen zou om het even welk cijfer lager dan 0.8 verder onderzocht moeten worden, want dit suggereert dat de beoordelaars niet op dezelfde manier te werk zijn gegaan.

Een indicatie van de betrouwbaarheid, zoals de Alpha die door MicroCAT (zie Tools voor statistische analyse hieronder) wordt geproduceerd, kan berekend worden voor de gehele groep van beoordelaars. Data kunnen klaargezet worden zoals beschreven in Figuur 18, met een paar aanpassingen: elke rij kan worden gebruikt om de respons van een kandidaat op een taak weer te geven; de kolommen kunnen worden gebruikt voor de scores van de beoordelaars.

## Many-Facet Rasch Measurement (MFRM)

Een meer gesofisticeerde manier om het werk van beoordelaars te beoordelen is technieken gebruiken zoals MANY-FACET RASCH MEASUREMENT (MFRM). Het is een variant van Rasch-analyse. MFRM kan gedaan worden met software die Facets genoemd wordt (Linacre, 2009). De analyse meet de moeilijkheidsgraad van de taken en de vaardigheid van de kandidaten zoals bij de Rasch-analyse, maar kan ook de strengheid of de mildheid van de beoordelaars evalueren. Het kan bovendien eerlijkere scores voor kandidaten opleveren omdat de effecten van de strengheid of mildheid van de beoordelaars verwijderd kunnen worden.

Wordt er MFRM gebruikt, dan is het heel belangrijk dat de data linken tussen beoordelaars bevatten, maar ook tussen kandidaten, taken en andere facetten die worden gemeten. Sommige prestaties moeten bijvoorbeeld door meerdere beoordelaars worden beoordeeld om een link te creëren tussen beoordelaars. Sommige kandidaten moeten meer dan één toets maken om een link te leggen tussen taken. Als er afzonderlijke dataverzamelingen gevormd worden, is MFRM niet in staat om schattingen te maken voor alle elementen.

Voor MFRM:

**minimaal aantal prestaties**: 30 voor elke taak die moet worden beoordeeld (Linacre, 2009)

**minimaal aantal beoordelingen door elke beoordelaar**: 30 (Linacre, 2009)

**meer informatie**: Eckes (2009)

# Constructvalidering

## De toetsstructuur controleren

Factoranalyse of *Structural Equation Modelling* kunnen helpen om na te gaan of de items van de toets het beoogde construct weergeven. De toets zou een patroon moeten reflecteren dat geïdentificeerd kan worden als het model voor taalgebruik dat werd aangenomen (zie hoofdstuk 1.1). Factoranalyse is erg nuttig in de fases van de toetsontwikkeling omdat het het meest gebruikt wordt om te controleren of de toets of de TOETSSPECIFICATIES werken zoals verwacht.

Voor factoranalyse:

**minimaal aantal kandidaten:** 200 (Jones, Smith en Talley, 2006, p. 495)

**meer informatie:** Verhelst (2004c)

## Item bias vaststellen

Item bias ontstaat wanneer een item op een onrechtvaardige manier een groep kandidaten bevoordeelt of benadeelt ten opzichte van kandidaten met dezelfde vaardigheden. Een item kan bijvoorbeeld makkelijker zijn voor vrouwelijke kandidaten, ook al beschikken de vrouwelijke en mannelijke kandidaten die de toets afleggen over dezelfde vaardigheden. Dit is onrechtvaardig want de bedoeling van de toets is dat die verschillen in taalvaardigheid meet, niet gender (zie hoofdstuk 1.4).

Er moet omzichtig worden omgesprongen met het opsporen van bias, want niet alle verschillen tussen groepen zijn onrechtvaardig. Leerders met een bepaalde L1 kunnen een item moeilijker vinden dan leerders uit een andere groep die over dezelfde vaardigheden beschikken omwillen van verschillen tussen de moedertaal en de doeltaal. In de context van taalvaardigheid meten, moet dit aanvaard worden als een deel van de eigenheid van de vaardigheid in de doeltaal en niet als een probleem om dit te meten.

Een aanpak om bias te beperken is gebruik maken van *Differential Item Functioning* (DIF) methodologie om mogelijke bias te detecteren zodat het verder kan onderzocht worden. Dit houdt in dat de responsen vergeleken worden van groepen kandidaten die over een vergelijkbare vaardigheid beschikken. Als de toets bijvoorbeeld gemaakt is voor volwassenen van alle leeftijden, dan zou de prestatie van jongere en oudere volwassenen met ongeveer dezelfde vaardigheden (volgens de toets) kunnen worden vergeleken. Analyses gebaseerd op IRT zijn erg geschikt om dit te doen.

Voor DIF-analyse met Rasch-analyse:

**minimaal aantal kandidaten:** 500, met tenminste 100 per groep (Jones, Smith en Talley, 2006, p. 495)

**meer informatie:** Camilli en Shepard (1994); Clauser en Mazor (1998)

## Steekproeven van kandidaten controleren

Als toetsdata worden gebruikt voor om het even welke analyse of onderzoek, dan moeten die representatief zijn voor de doelgroep van kandidaten (de populatie). Informatie over deze kandidaten kan regelmatig verzameld en gecontroleerd worden om te zien of de analyse werd uitgevoerd met een representatieve steekproef van kandidaten.

Demografische gegevens over kandidaten kunnen verzameld worden telkens als er een toets wordt afgenomen (zie hoofdstuk 4). Kenmerken kunnen vergeleken worden door simpele percentages te gebruiken, bijvoorbeeld het evenwicht tussen mannen en vrouwen in twee verschillende steekproeven.

Een meer gesofisticeerde analyse zal ook proberen te bepalen of er verschillen tussen steekproeven bestaan die misschien toevallig zijn. Een Chi-square toets kan op die manier gebruikt worden. De resultaten van een analyse moeten dan kwalitatief gecontroleerd worden om te zien of er verschillen zijn die aanleiding kunnen geven tot substantieve verschillen in de prestaties van de kandidaten.

## Tools voor statistische analyses

Er bestaan een aantal commerciële softwarepakketten die geschikt zijn voor dit soort werk. Sommige berekeningen kunnen redelijk eenvoudig uitgevoerd worden in Microsoft Excel of in andere veelgebruikte spreadsheat programma's. Gespecialiseerde aanbieders worden in alfabetische volgorde opgelijst en kunnen tools bieden voor verschillende soorten analyses. Studenten- of demoversies zijn soms beschikbaar.

| | |
|---|---|
| Assessment Systems | http://www.assess.com/softwarebooks.php |
| Curtin University of Technology | http://lertap.curtin.edu.au/index.htm |
| RUMM Laboratory | http://www.rummlab.com.au/ |
| Winsteps | http://www.winsteps.com/index.htm |

Andere gratis tools zijn beschikbaar voor specifieke doeleinden:

| | |
|---|---|
| William Bonk, University of Colorado | http://psych.colorado.edu/~bonk/ |
| Del Siegle, University of Connecticut | http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability/reliabilitycalculator2.xls |

# Bijlage VIII – Verklarende woordenlijst

**actiegerichte aanpak**

Een manier van nadenken over taalvaardigheid waarbij taal wordt gezien als een middel om in een sociale context communicatieve 'acties' uit te voeren.

**afname**

De datum waarop of de periode waarin een toets plaatsvindt. Veel toetsen worden meerdere malen per jaar op een vaste datum afgenomen, andere kunnen op aanvraag worden afgenomen.

**ankeritem**

Een item dat in twee of meer toetsen voorkomt. De kenmerken van ankeritems zijn bekend. Ankeritems maken deel uit van een nieuwe versie van een toets en worden gebruikt om informatie te krijgen over deze toets en de kandidaten die de toets hebben afgelegd, dat wil zeggen om een nieuwe toets op een MEETSCHAAL te kunnen kalibreren.

**authenticiteit**

De mate waarin de taken van de toets een afspiegeling zijn van taalgebruik in een levensechte, niet-toetsgebonden situatie. Bijvoorbeeld notities nemen tijdens een taaltoets voor academisch taalgebruik, in plaats van alleen maar te luisteren. Zie ook het *nut van een toets*.

**belanghebbenden**

Mensen en organisaties die belang hebben bij de toets zoals kandidaten, scholen, ouders, werkgevers, regeringen en werknemers.

**beoordelaar**

Iemand die een score toekent aan de prestatie van een kandidaat op een toets en daarbij een subjectieve beoordeling hanteert. Beoordelaars zijn meestal vakdocenten die extra instructie krijgen (bijvoorbeeld in het normeren). Bij mondelinge toetsing wordt er soms een verschil gemaakt tussen degene die het gesprek leidt/de gesprekspartner en degene die beoordeelt.

**beoordelingsschaal**

Een schaal bestaande uit een aantal geordende categorieën voor het geven van een subjectieve beoordeling. In taaltoetsing worden de beoordelingsschalen meestal aangeboden met descriptoren die aangeven hoe deze schalen moeten worden geïnterpreteerd.

**bereik**

Een maat van spreiding van waarnemingen. Het bereik is de afstand tussen de hoogste en de laagste waarneming.

**betrouwbaarheid**

De consistentie of stabiliteit van de maten die een toets heeft voortgebracht. Hoe betrouwbaarder de toets, hoe minder random meetfout deze bevat. Een toets met systematische meetfout, bijvoorbeeld bias ten opzichte van een bepaalde groep, kan wel betrouwbaar zijn, maar is niet valide.

**cesuurbepaling**

Het proces waarbij een cesuur (de grens tussen slagen of niet slagen) wordt bepaald voor een toets en dus de betekenis van de toetsresultaten.

**construct**

Een hypothetisch vermogen of een mentale EIGENSCHAP die zich niet direct laat observeren of meten, zoals bijvoorbeeld in een taaltoets luistervaardigheid.

**construct-irrelevante variantie**[7]

Meetfout.

**correctieschema**

Een lijst met alle acceptabele antwoorden op de items in een toets. Met een correctieschema is een corrector in staat een toets heel precies na te kijken.

**corrector**

Iemand die de responsen van een kandidaat nakijkt. Het kan hierbij gaan om iemand met kennis van zaken of, bij methodische correctie, om het toepassen van een correctieschema, zonder specifieke kennis van zaken.

**correlatie**

De sterkte van een relatie tussen twee of meer metingen: in hoeverre variëren ze op dezelfde manier? Als kandidaten bijvoorbeeld op twee verschillende toetsten op vergelijkbare manier worden geordend, dan is er sprake van positieve correlatie tussen de twee scoresets.

7 Noot van de vertaalster: voor een goed begrip van de tekst toegevoegd aan de vertaling.

**descriptor**

Een korte beschrijving bij een band op een beoordelingsschaal: een korte aanduiding van de mate van kennis van een taal en de bedrevenheid in het gebruik ervan of het soort prestatie dat van een kandidaat bij die score verwacht wordt.

**dichotoom item**

Een item dat goed of fout kan worden gerekend. Dichotome items omvatten bijvoorbeeld meerkeuze, goed/fout en items waarvoor een kort antwoord wordt verwacht.

**discreet item**

Een op zichzelf staand item. Hoort niet bij andere items of bij aanvullend materiaal.

**discriminatie**

Het vermogen van een item om een onderscheid te maken tussen zwakkere en sterkere kandidaten. Er worden verschillend discriminatie-indexen gebruikt. Zie Bijlage VII voor meer informatie.

**domein van taalgebruik**

Brede domeinen van het sociale leven zoals onderwijs of privéleven, die kunnen worden afgebakend voor de selectie van inhoud en vaardigheden voor een toets.

**doorlichten**

Een stadium in het samenstellen van een toets waarbij de toetsontwikkelaars materiaal van itemschrijvers beoordelen en bepalen welke items moeten worden afgewezen omdat ze niet voldoen aan de toetstoetsspecificaties en welke door kunnen naar het redigeerstadium.

**dubbele beoordeling**

Een methode voor het evalueren van prestaties waarbij twee personen onafhankelijk van elkaar de prestaties van een kandidaat op een toets beoordelen.

**equivalente vormen**

Dit worden ook wel *parallelle* of *alternatieve vormen* genoemd. Verschillende versies van dezelfde toets die als equivalent worden gezien omdat ze dezelfde toetsspecificaties als basis hebben en dezelfde vaardigheid meten. Om binnen de klassieke toetstheorie aan de strenge equivalentie-eisen te kunnen voldoen, moeten verschillende vormen van een toets dezelfde gemiddelde moeilijkheidsgraad, variantie en covariantie met een concurrent criterium vertonen wanneer ze aan dezelfde personen worden voorgelegd. Equivalentie is in de praktijk moeilijk te verwezenlijken.

**gebruiksbewijs**

Dat deel van het validiteitsbewijs dat verantwoordt hoe de resultaten van de toets geïnterpreteerd zouden moeten worden als een toets voor een bepaalde context wordt gebruikt.

**gemiddelde**

Centrale-tendentiemaat. Een gemiddelde toetsscore berekent men door alle scores op de toets op te tellen en de som te delen door het aantal scores.

**high stakes**[8]

*High stakes* staat tegenover *low stakes*. High stakes toetsen hebben veel impact op de kandidaat.

**impact**

De mate waarin de resultaten van een toets de toekomst van een kandidaat, maar ook de maatschappij en educatieve processen kunnen beïnvloeden.

**input**

Materiaal in een toetstaak dat een kandidaat kan gebruiken om de juiste respons te leveren. In een luistertoets kan dit bijvoorbeeld een audio-opname van een tekst zijn met een aantal schriftelijke items.

**instructies**

De instructies die de kandidaten ontvangen om hun antwoorden op een toetstaak in bepaalde banen te leiden.

**interactiviteit**

De mate waarin items en toetstaken mentale processen en strategieën vragen die ook in levensechte taken verwacht worden. Zie ook het *nut van een toets*.

**interpretatief bewijs**

zie *gebruiksbewijs*

**intervalschaal**

Een meetschaal waarop de afstand tussen twee aan elkaar grenzende meeteenheden gelijk is, maar een betekenisvol nulpunt ontbreekt.

**item**

Een onderdeel van een toets waaraan een apart cijfer wordt gegeven. Voorbeelden: een gat in een cloze-toets, een meerkeuzevraag met drie of vier keuzemogelijkheden, een zin die grammaticaal

8 Noot van de vertaalster: voor een goed begrip van de tekst toegevoegd aan deze Verklarende woordenlijst.

moet worden omgezet, een vraag waarop met een zin moet worden geantwoord.

**itemanalyse**

Een beschrijving van het functioneren van individuele toetsitems, meestal door gebruik te maken van beschrijvende statistieken zoals proportie correct en discriminatie. Bij deze analyse wordt gebruik gemaakt van software zoals bijvoorbeeld MicroCAT.

**itembank**

Een methode voor het beheren van toetsitems, dat wil zeggen het opslaan van informatie over items zodat er toetsen met een van tevoren bekende inhoud en moeilijkheidsgraad kunnen worden opgesteld.

**itemresponstheorie (IRT)**

Een groep mathematische modellen voor het leggen van verbanden tussen iemands prestatie op een toets en het vaardigheidsniveau van deze persoon. De modellen gaan uit van de theorie dat de te verwachten prestatie op een bepaalde toetsvraag of item te maken heeft met zowel de moeilijkheidsgraad van het item als het vaardigheidsniveau van de kandidaat.

**kalibratie**

Het bepalen van de schaal van één of meerdere toetsen. Tijdens een kalibratie kunnen items uit verschillende toetsen worden geankerd op een gemeenschappelijke moeilijkheidsschaal. Als een toets bestaat uit gekalibreerde items, dan geven de toetsscores een indicatie van de vaardigheid van een kandidaat, dat wil zeggen hun plaats op de schaal.

**kalibreren**

In de itemresponstheorie, het inschatten van de moeilijkheidsgraad van een set van toetsitems.

**linken**

Een procedure om resultaten van (een vorm van) een toets te vertalen naar de resultaten van een andere toets of vorm van een toets. Deze procedure helpt om verschillen in toetsmoeilijkheid of vaardigheden van kandidaten te compenseren.

**live toets**

Een toets die afgenomen wordt en die om die reden geheim moet blijven.

**logit**

De meeteenheid gebruikt in de IRT/Rasch-analyse en MFRM.

**Many-Facet Rasch Measurement (MFRM)**

Een uitbreiding van het Rasch-model. De moeilijkheidsgraad van de items of de vaardigheid van de kandidaten wordt in facetten uit elkaar gehaald zodat de data die gelinkt zijn aan deze afzonderlijke facetten gebruikt kunnen worden om de scores uit te leggen die aan elke kandidaat worden toegekend. Zo kan de strengheid van de beoordelaar helpen om de score uit te leggen die de kandidaat gekregen heeft voor een schrijftaak. In dit geval worden scores gezien als het geheel van vaardigheden van de kandidaat, de moeilijkheidsgraad van de taak en de strengheid van de beoordelaar. Het is dan mogelijk om de impact van de strengheid van de beoordelaar uit te sluiten van de definitieve score die aan de kandidaat wordt toegekend.

**matching-taak**

Een soort toetstaak waarbij elementen uit twee verschillende lijsten moeten worden gecombineerd. Een voorbeeld van een matching-toets is een lijst met zinnen die niet af zijn. Uit een tweede lijst moeten vervolgens de zinsdelen worden gekozen waarmee de zinnen kunnen worden aangevuld. In leesvaardigheidstoetsen wordt bijvoorbeeld een beschrijving van een persoon gegeven en moet in een lijst worden opgezocht welke vakantie of welk boek bij deze persoon passen.

**meetfout**

De meetfout, ook wel de standaard meetfout genoemd, is een indicatie voor de onnauwkeurigheid van een meting. Als de meetfout bijvoorbeeld 2 is, dan zal een kandidaat met een score van 15 (met 68% zekerheid) een score tussen 13 en 17 hebben. Een kleinere fout zal leiden tot een meer precieze score. Zie ook *construct-irrelevante variantie*.

**meetschaal**

Een meetschaal is een schaal bestaande uit cijfers die kan gebruikt worden om het verschil te meten tussen kandidaten, items, cesuren, … in relatie tot het toetsconstruct. Een meetschaal ontstaat door statistische technieken toe te passen op de antwoorden van kandidaten op items (zie Bijlage VII). Meetschalen geven meer informatie dan ruwe scores omdat ze bijvoorbeeld niet alleen tonen welke kandidaten beter zijn dan andere, maar ook hoeveel beter ze zijn. Soms wordt er met meetschaal verwezen naar nominale en ordinale schalen, maar deze definitie is niet gebruikt in deze handleiding.

**methodische correctie**

Een corrigeermethode waarbij de correctoren geen speciale expertise nodig hebben en geen eigen oordeel hoeven te geven. Ze corrigeren aan de hand van een correctieschema waarin alle acceptabele antwoorden op een toetsitem staan weergegeven.

**model fit**

Als een model (zoals het Rasch-model) in statistische analyses wordt gebruikt, dan is het belangrijk om te bekijken hoe goed het model bij de data past. Een model staat voor de ideale weergave van de data, dus een perfecte model fit is niet aan de orde. Toch kan een hoge misfit betekenen dat de conclusies die over de data getrokken worden, niet geldig zijn.

**model van taalgebruik**

Een beschrijving van de vaardigheden die nodig zijn voor taalgebruik en de manier waarop ze met elkaar zijn verbonden. Een model is een basiscomponent van een toetsontwerp.

**nut van een toets**

Het nut van een toets (Bachman en Palmer, 1996) is het idee dat een toets het meest zinvol is als het evenwicht tussen de validiteit, betrouwbaarheid, authenticiteit, interactiviteit, impact en praktische haalbaarheid optimaal is.

**objectieve correctie**

Items waaraan een score kan toegekend worden aan de hand van een correctieschema, zonder dat er nood is aan het oordeel van deskundigen of een subjectieve beoordeling.

**open taak**

Een soort item of toetstaak in een schriftelijke toets waarop de kandidaat met actieve productie moet antwoorden en er niet kan gekozen worden uit een aantal mogelijke antwoorden. Dit soort items wil een relatief vrij antwoord uitlokken, dat in lengte kan verschillen, gaande van een paar woorden tot een uitgebreid essay. Het correctieschema laat daarom de ruimte voor meerdere juiste antwoorden.

**op tekst gebaseerd item**

Een item gebaseerd op een lopende tekst, bijvoorbeeld meerkeuze-items gebaseerd op een leestekst.

**optische lezer**

Een elektronisch apparaat voor het scannen van antwoordbladen. Kandidaten of beoordelaars kunnen de juiste antwoorden op een antwoordblad aangeven. Deze informatie wordt vervolgens door de computer gelezen. Wordt ook wel scanner genoemd.

**partial credit item**

Een item dat niet noodzakelijk volledig goed moet beantwoord worden om er punten voor te krijgen. De score die aan een antwoord wordt toegekend kan bijvoorbeeld variëren van 1, 2 of 3 naargelang het niveau van correctheid dat in de sleutel wordt beschreven.

**pilotstudie**

Een pilotstudie (of pilot) is een studie vooraf, waarin onderzoekers of toetsontwikkelaars hun ideeën op een beperkte groep uitproberen. Collega's kunnen bijvoorbeeld gevraagd worden om items te beantwoorden en feedback te geven.

**praktische haalbaarheid**

De mate waarin, rekening houdende met de beschikbare middelen, een toets die aan vooropgestelde eisen voldoet ontwikkeld kan worden . Zie ook het *nut van een toets*.

**pretesten**

Een stadium in de ontwikkeling van toetsmaterialen waarin de items worden uitgeprobeerd bij representatieve steekproeven van de doelpopulatie om de moeilijkheidsgraad van deze items te bepalen. Na statische analyse kunnen de items die goed zijn bevonden gebruikt worden in de definitieve versie van de toets.

**proefafname**

Een stadium in de ontwikkeling van toetstaken dat tot doel heeft te bepalen of de toets naar verwachting functioneert. Gaat vaak gepaard met subjectief te beoordelen taken, bijvoorbeeld essayvragen die aan een beperkte populatie worden voorgelegd.

**proportie correct**

De proportie juiste antwoorden op een item, uitgedrukt als een waarde tussen 0 tot 1. Wordt soms ook weergegeven als percentage. Wordt ook wel faciliteitswaarde, gemakkelijkheidswaarde of p-waarde genoemd.

**Rasch-analyse**

Een analyse gebaseerd op een mathematisch model, ook gekend als het *simple logistic model*. Het Rasch-model gaat uit van een relatie tussen de mate van waarschijnlijkheid waarmee iemand een taak tot een succesvol einde zal brengen en het verschil tussen de vaardigheid van de persoon in kwestie en de moeilijkheidsgraad van de taak.

Het is de wiskundige equivalent van het een-parametermodel in de itemsresponstheorie.

**register**

Een specifieke manier van spreken of schrijven die karakteristiek is voor een bepaalde activiteit of een bepaalde mate van formaliteit.

**respons**

Het gedrag van een kandidaat opgeroepen door de toetsinput. Bijvoorbeeld het antwoord op een multiplechoicevraag of het werk dat door een kandidaat wordt geproduceerd in een schrijftaak. Ook wel prestatie genoemd, al wordt deze term doorgaans gebruikt om te verwijzen naar het geheel van responsen geproduceerd door een kandidaat die een toets aflegt.[9]

**responsstimulus**

Grafisch materiaal of teksten bij spreek-of schrijftoetsen, ontworpen om een bepaalde respons van de kandidaat uit te lokken.

**ruwe score**

Een toetsscore die niet statistisch gemanipuleerd is door transformatie, weging of herschaling.

**samenstellen van een toets**

Het samenstellen van een toets gebeurt door items of taken uit te kiezen. Dit proces wordt dikwijls voorafgegaan door een stadium waarin materialen worden gepretest of aan een proefafname onderworpen. Items en taken voor een toets kunnen ook uit een itembank worden gehaald.

**schaal**

Een aantal getallen of categorieën waarmee iets kan worden gemeten. Er worden vier soorten meetschalen onderscheiden: nominale, ordinale, interval en ratioschalen.

**schrijfproduct**

De schriftelijke antwoorden van kandidaten op een toets, met name op een toets met open vragen.

**sleutel**

a) De correcte oplossing bij een meerkeuze-item.

b) Meer algemeen: alle correcte of acceptabele antwoorden op toetsitems samen.

**standaarddeviatie (SD)**

Een maat voor de spreiding van toetsscores (of andere data) rond het rekenkundig gemiddelde. Bij een normale verdeling van de scores bevindt 68% van de scores zich binnen één standaarddeviatie van het gemiddelde en 95% binnen twee standaarddeviaties van het gemiddelde. Hoe hoger de standaarddeviatie, hoe groter de spreiding.

**subjectieve correctie**

Items waaraan een score toegekend wordt op basis van het oordeel van deskundigen of een subjectieve beoordeling.

**taak**

Wat een kandidaat wordt gevraagd om te doen om een deel van een toets af te leggen. Een taak uitvoeren is complexer dan één enkel, op zichzelf staand item beantwoorden. Een taak refereert meestal aan een mondelinge of een schriftelijke prestatie of aan een serie van items die op de een of andere manier met elkaar verbonden zijn. Een leestekst kan bijvoorbeeld gepaard gaan met verschillende meerkeuzevragen. De kandidaat kan deze vragen beantwoorden aan de hand van een en dezelfde instructie.

**toetsonderdeel**

Deel van een toets, meestal aangeboden als aparte toets, met eigen instructieboekje en tijdslimiet. Toetsonderdelen zijn vaak gebaseerd op een specifieke vaardigheid. Zo is er bijvoorbeeld een onderdeel Luistervaardigheid of Steltaak.

**toetsontwikkelaar**

Iemand die zich bezighoudt met het ontwikkelen van een toets.

**toetsspecificaties**

Een lijst met gedetailleerde informatie over een toets, zoals wat er wordt getoetst, hoe dit gebeurt, het aantal en de lengte van de schriftelijke antwoorden, soorten items die worden gebruikt,…

**toezichthouder**

Persoon die tijdens de toetsafname, in de ruimte waar de toets wordt afgenomen, de eindverantwoordelijkheid draagt.

**validering**

Het verzamelen van gegevens ter ondersteuning van de conclusies uit toetsscores.

**validiteit**

De mate waarin conclusies op basis van toetsscores juist en zinvol zijn, gegeven het doel van de toets.

**validiteitsbewijs**

Een uitgebreide reeks van beweringen en bijhorende gegevens dat de validiteit van de gegeven interpretaties van de toetsresultaten onderbouwt.

9 Noot van de vertaalster: Voor een goed begrip van de tekst toegevoegd.

**vraag**

Wordt soms gebruikt om te verwijzen naar een 'toetstaak' of 'item'.

**waarderen**

Het proces waarbij toetsscores in een waardering worden vertaald.

**waardering**

Een toetsscore kan aan de kandidaat worden meegedeeld in de vorm van een waardering, bijvoorbeeld op een schaal van A tot E, waarbij A staat voor 'zeer goed', B voor 'goed', C voor 'voldoende' en D en E voor 'onvoldoende'.

**wegen**

Het toekennen van verschillende maximumscores aan verschillende toetsitems, -taken, of -onderdelen om zo de relatieve bijdrage te wijzigen. Als er bijvoorbeeld een dubbele score wordt toegekend aan alle items in taak 1 van een toets, dan zal taak 1 meer van de totaalscore voor zijn rekening nemen dan de andere taken.

**Deze verklarende woordenlijst is gebaseerd op de *Multilingual Glossary of Language Testing Terms* geproduceerd door de Association of Language Testers in Europe (ALTE-leden 1998) en de *Dictionary of Language Testing* (Davies et al 1999), beide gepubliceerd door Cambridge University Press in de *Studies of Language Testing* serie. De aanvullende ingangen werden geschreven zoals vereist.**

# Dankwoord

Deze handleiding is een gereviseerde versie van een handleiding die eerder gepubliceerd werd door de Raad van Europa in 2002 onder de titel *Language Examination and Test Development*. Dat document was een versie van de *Users's Guide for Examiners*, gepubliceerd door ALTE in opdracht van de Raad van Europa in 1996.

**De Raad van Europa wil graag haar dank betuigen aan**

de Association of Language Testers in Europe (ALTE) voor het reviseren van dit document.

**de redacteurs van deze herziene versie:**

David Corkill
Michael Corrigan
Neil Jones
Michael Milanovic
MartinNuttall
Nick Saville

**de leden van de ALTE CEFR/Manual Special Interest Group en hun collega's voor het verstrekken van aanvullend materiaal en het reviseren van de verschillende kladversies van de tekst:**

Elena Archbold-Bacalis
Sharon Ashton
Andrew Balch
Hugh Bateman
Lyan Bekkers
Nick Beresford-Knox
Cris Betts
Margherita Bianchi
Inmaculada Borrego
Jasminka Buljan Culey
Cecilie Carlsen
Lucy Chambers
Denise Clarke
Maria Cuquejo
Emyr Davies
Desislava Dimtrova
Angela ffrench
Colin Finnerty
Anne Gallagher
Jon-Simon Gartzia
Annie Giannakopoulou
Begona Gonzales Rei
Guiliana Grigorova
Milena Grigorova
Ines Hälbig
Berit Halvorsen
Marita Harmala
Martina Hulešová
Nurita Jornet
Marion Kavallieros
Gabriele Kecker
Kevin Kempe
Wassilios Klein
Mara Kokina
Zsofia Korody
Henk Kuijper
Gad Lim
Juvana Llorian
Karen Lund
Lucia Luyten
Hugh Moss
Tatiana Nesterova
Desmond Nicholson
Gitte Østergaard Nielsen
Irene Paplouca
Szilvia Papp
Francesca Patrizzi
Jose Ramón Rarrondo
Jose Pascoal
Roberto Perez
Michaela Perlmann-Balme
Tatiana Perova
Sibylle Plassmann
Laura Puigdomenech
Meilute Ramoniene
Lýdia Rihová
Shelagh Rixon
Martin Robinson
Lorenzo Rocca
Shalini Roppe
Dittany Rose
Angeliki Salamoura
Lisbeth Salomonsen
Gergio Silfer
Gabriela Šnaidaufová
Ioana Sonea
Annika Spolin
Stefanie Steiner
Michaela Stoffers
Gunlog Sundberg
Lynda Taylor
Julia Todorinova
Rønnaug Katharina Totland
Gerald Tucker
Piet Van Avermaet
Mart van der Zanden
Elorza Juliet Wilson
Beate Zeidler
Ron Zeronis

**de revisoren van de Raad van Europa:**

Neus Figueras
Brian North
Johanna Panthier
Sauli Takala

**het uitgeversteam:**

Rachel Rudge
Gary White

**Deze vertaling is een product van:**

de vertaalster:
Christina Maes

de revisoren:
Bart Deygers
Inge Reinders
Lies Strobbe

de eindredacteur:
Steven Verheyen

de vormgevers:
Jurgen Leemans (tekst)
Sophie Willems (cover)

De Association of Language Testers in Europe (ALTE), met haar status van ondersteunende Internationale Niet-Gouvernementele Organisatie (INGO) binnen de Raad van Europa, heeft bijgedragen tot de bronnen waaruit de toolkit van de Raad van Europa is opgebouwd, zoals de EAQUALS/ALTE European Language Portfolio (ELP) en de CEFR Grids for Analysis of speaking and writing tasks.

Samen met de Language Policy Division van de Raad van Europa, moedigt ALTE de gebruikers van de toolkit aan om het ERK efficiënt te gebruiken in hun specifieke context om hun eigen doelstellingen te bereiken.

Deze vertaling is een initiatief van:

Certificaat Nederlands als Vreemde Taal (CNaVT)
p/a Centrum voor Taal en Onderwijs (KU Leuven)
Blijde-Inkomststraat 7 bus 3319
BE-3000 Leuven
België
www.cnavt.org

Met toestemming van de Raad van Europa.

*De Raad van Europa is de toonaangevende mensenrechtenorganisatie van Europa. De Raad telt 47 lidstaten, waarvan 28 ook lid zijn van de Europese Unie. Alle lidstaten van de Raad van Europa hebben het Europees Verdrag voor de Rechten van de Mens ondertekend. Dit Verdrag is gericht op de bescherming van mensenrechten, democratische waarden en rechtsstaatbeginselen. Het toezicht op de uitvoering van het Verdrag in de lidstaten is in handen van het Europees Hof voor de Rechten van de Mens.*

ALTE Association of Language Testers in Europe

taal: unie CNaVT Certificaat Nederlands als Vreemde Taal

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE