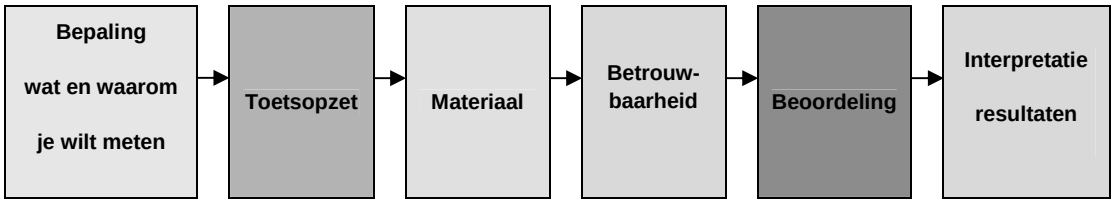

TIP 9: BEOORDELINGSMODELLEN: PRAKTISCH

In de vorige toetstip kon u lezen over beoordelingsmodellen: waar dienen ze precies voor? Hoe zien ze eruit? Welke soorten bestaan er en wat zijn dan de voor- en nadelen?
Deze toetstip begeleidt u als u zelf een beoordelingsmodel wilt ontwikkelen: waarom zou u dat willen doen en hoe pakt u dat dan het best aan?
Gesloten toetsvragen (zoals meerkeuzevragen, matchoefeningen,…) vragen geen ingewikkeld beoordelingsmodel. Open taken vragen dat wel. Deze toetstip focust zich dan ook op beoordelingsmodellen bij open vragen.

Er zijn veel redenen om met een beoordelingsmodel te gaan werken:

- U wilt niet 'uit de losse pols' een score op een toetstaak plakken.
- U wilt stilstaan bij uw huidige manier van beoordelen.
- U wilt studenten onderbouwde feedback kunnen geven, op basis van objectief aantoonbare beoordelingsaspecten.
- Het is de gewoonte in uw onderwijsinstelling dat wie meer dan 50% haalt, geslaagd is. U stelt zich daar vragen bij en wilt op basis van een beoordelingsmodel dit beleid herzien.
- U ontwikkelt een toets die een grote impact heeft op het leven van een kandidaat (bv. een toelatingsexamen) en wilt dat grondig aanpakken.
- …

Welke van bovenstaande redenen u ook voor een beoordelingsmodel doen kiezen, het zijn allemaal aanwijzingen dat u een betrouwbare toets wilt maken.

Ter herinnering, een toets is betrouwbaar als je met een gerust hart conclusies kunt trekken op basis van het toetsresultaat van een kandidaat: "Hoe kunt u conclusies trekken op basis van een score als u niet zeker bent van die score?". Een betrouwbare toets vereist een consistente beoordeling. Concreet betekent dit twee dingen:

- Als test *w*, afgelegd door student *x*, tweemaal verbeterd wordt door docent *y*, krijgt student *x* tweemaal dezelfde score.

- Als test *w*, afgelegd door student *x*, verbeterd wordt door docent *y* en docent *z*, krijgt student *x* tweemaal dezelfde score.

Een beoordelingsmodel is een waardevol hulpmiddel om een betrouwbare toets te maken. Wanneer je als toetsconstructeur gaat bepalen op welke criteria toetsprestaties beoordeeld worden, ben je verplicht om objectief te beoordelen.

Beoordelingsmodellen zijn echter niet per definitie en zonder meer waardevol; wanneer er niet goed over nagedacht is, betekent het weinig voor de betrouwbaarheid. In sommige gevallen kan het zelfs het tegengestelde effect hebben. Het beoordelingsmodel vertaalt de prestatie van de kandidaat dan naar een score die niet overeenstemt met de werkelijke vaardigheid van de kandidaat.

Zoals gezegd is het niet altijd makkelijk om bij open opdrachten een beoordelingsmodel te ontwikkelen. Kijk bijvoorbeeld naar de volgende opdracht:

Vul aan:

Na de les ga ik ....

Mag ik....

Zelfs voor een eenvoudige aanvuloefening als deze is het niet gemakkelijk om een betrouwbaar beoordelingsmodel te maken. Een voor de hand liggende mogelijkheid zou zijn:

De kandidaat vult aan tot een correcte zin: 1

De kandidaat vult niet aan tot een correcte zin: 0

Maar stel nu dat een kandidaat de volgende oplossing geeft:

Vul aan:

Na de les ga ik goed studeren, zodat ik mijn volgende toets goed aflech.

Mag ik gaan?

Voor de tweede zin scoort hij dan een 1, hoewel hij weinig taalvaardigheid heeft laten zien. De eerste zin is op dat gebied een stuk beter, maar wat met de spellingsfout? Krijgt de kandidaat daardoor een 0?

De vragen die u zich moet stellen wanneer u een beoordelingsmodel ontwikkelt, zijn: Waarom? Wat? Hoe? Deze vragen komen hieronder uitgebreid aan bod.

WAAROM?

Waarom neem ik deze toets af?

Twee voor de hand liggende antwoorden op deze vraag zijn: om te beslissen of een kandidaat al dan niet geslaagd is voor een examen, of om meer te weten te komen over de soorten fouten die de kandidaat nog maakt. In dat tweede geval heeft de toets een diagnostische functie en is een meer gedetailleerd beoordelingsmodel nodig om de kandidaat die informatie te kunnen geven die het hem mogelijk maakt om vorderingen te maken. Naar gelang het doel van de toets is een ander soort beoordelingsmodel wenselijk.

Ook de grootte van de getoetste groep speelt mee bij de beslissing om met een (uitgebreid) beoordelingsmodel te werken: bij een ingangsexamen bijvoorbeeld, waarbij een groep kandidaten onderverdeeld wordt in bv. geslaagd of niet geslaagd, gaat het meestal om grote groepen kandidaten en zal de beoordeling zeer efficiënt moeten verlopen. Hier streeft de ontwikkelaar het best naar een minder gedetailleerd beoordelingsmodel. Een klas studenten die gedetailleerde feedback verwachten over waar ze precies staan in hun taalleerproces, is meestal een veel kleinere groep. Afhankelijk van het doel van de toets, kan de ontwikkelaar het beoordelingsmodel zo gedetailleerd als wenselijk en nodig uitwerken.

Richtvragen bij de keuze van het soort beoordelingsmodel:

- Hoe groot is de groep en wat is het belang van efficiëntie?
- Hoe gedetailleerd moet de eventuele feedback zijn?
- Hoeveel tijd is er om de toets te ontwikkelen? En om de prestaties te verbeteren? Er zijn modellen (of modellen voor deelaspecten) die je voor elke taak kunt gebruiken (die zijn grof en het kan enkel bij open taken) en je hebt modellen voor specifieke taken. Welke soort is optimaal voor mijn opzet?
- Verder is het mogelijk om te beoordelen met cijfers (bv. 6/10) of met een waardeoordeel (voldoende, zeer goed,...) Moet het eindresultaat een concreet cijfer zijn? Of moet het een uitspraak over het niveau van de prestatie zijn?

De volgende stelregels kunnen helpen:

Wordt de toets ingezet voor de beoordeling van een **grote groep**, dan is een generisch of analytisch[1] model een goede optie, omdat het een snelle beoordeling mogelijk maakt.

Om een **kleine groep** kandidaten gedetailleerde feedback te geven, gaat u beter aan de slag met een primary trait of analytisch model.

Ook de haalbaarheid speelt een rol: hebt u weinig tijd, kies dan voor een generisch of beperkt analytisch model, is er voldoende tijd, dan kunt u een uitgebreid analytisch of een primary trait-model ontwikkelen. In veel gevallen blijken analytische modellen een goede middenweg te vormen.

De manier waarop de beoordelingsmodellen verder uitgewerkt worden, hangt samen met de vragen 'Wat' en 'hoe' toetsen we?

WAT?

Welke informatie over de taalvaardigheid van de leerder hebben we nodig?

Wat moet de toets precies meten? Leesvaardigheid? Schrijfvaardigheid? Gaat het om de kwaliteit? Om de kwantiteit? Om beide? Een goed beoordelingsmodel weerspiegelt wat u precies wilt meten. In deze fase beslist u wat de criteria zullen zijn waarop u de prestaties wilt beoordelen.

Om de beoordelingscriteria te bepalen is het goed terug te keren naar de vraag 'Wat toets ik'? Een (deel)vaardigheid, bvb. luisteren? Een specifieker segment van taalvaardigheid, bvb. notities maken bij een academische mondelinge uiteenzetting? Of moet een kandidaat zichzelf kort kunnen voorstellen en speelt vormcorrectheid daarbij geen grote rol?

Richtvragen:

- Welke (sub)vaardigheid wil ik meten?
- Primeert inhoud of vorm of zijn beide gelijkwaardig?
- Welke aspecten zijn nog van belang voor deze taak?
- …

Wanneer u een beoordelingsmodel ontwikkelt, kijk dan naar prestaties van studenten op een taak die lijkt op de taak waarvoor het nieuwe model ontwikkeld wordt. Dat kan u helpen om de volgende vragen te beantwoorden: *Wat moet de student minstens schrijven of zeggen om voldoende te scoren?*

[1] Meer uitleg bij de verschillende soorten beoordelingsmodellen vindt u in de vorige toetstip. Die kunt u vinden op www.CNaVT.org , doorklikken naar de toetstips.

Hoe kunt u scoren als een kandidaat onverwachte antwoorden geeft? Welke vormelijke eisen stelt u aan de producten van uw kandidaten?

En vooral: *Wat maakt dat dit een 'goede' prestatie is en een andere een 'minder goede'? Welke aspecten precies spelen hierbij een rol?*

Op basis van deze analyse kunt u de beoordelingscriteria destilleren die in het model moeten komen.

Laten we een voorbeeld bekijken:

Stel dat de kandidaten hun mening moeten geven over een stelling naar keuze. De opdracht luidt: "Kies een van deze drie stellingen. Bent u voor of tegen, of hebt u een genuanceerde mening? Onderbouw uw antwoord met drie verschillende argumenten."

- In een generisch model nemen we de criteria op: argumenteren de kandidaten en hoe komt dat globaal over?
- Een analytisch model zal meer gedetailleerde informatie opleveren op basis van verschillende criteria, bv. volledigheid van de inhoud, accuraatheid van de argumenten, spelling, grammatica en opbouw.
- Een primary trait-model zal nog dieper ingaan op de specifieke link tussen de stellingen in de opgave en de argumenten die de kandidaat daarbij geeft.

Tips voor de selectie van beoordelingscriteria:

Om met een gerust hart en een zuiver geweten de beoordelingscriteria te bepalen kunt u verschillende instrumenten raadplegen. U kunt uw beoordelingscriteria baseren op de doelstellingen uit uw curriculum, leerplan of leermethode, maar ook op de descriptoren uit het Europees Referentiekader (ERK). Op die manier is er een wisselwerking tussen toets(taak) en beoordelingsmodel, waarbij u nog meer solide gaat nadenken over de toetsinhouden en de vereisten voor de kandidaat.

Zorg ervoor dat er geen overlap is tussen de verschillende criteria:

Stel dat in het beoordelingsmodel het criterium 'de boodschap overbrengen' voorkomt, alsook de criteria 'opbouw' en 'structuur'. Bent u dan niet twee keer, of zelf drie keer hetzelfde aan het beoordelen? Een kandidaat die zijn boodschap geslaagd opbouwt en structureert, zal namelijk ook beter scoren op 'de boodschap overbrengen' als geheel. Veel hangt dus af van de afbakening en de precieze omschrijving van elk criterium.

- Zorg er ook voor dat de te beoordelen items onafhankelijk zijn. We verduidelijken dit met een voorbeeld: De kandidaat moet – op basis van een luisterfragment – een wegbeschrijving aanduiden op een plattegrond. Als de kandidaat al bij de eerste stap de verkeerde kant uit gaat, scoort hij dan op elke volgende stap ook een nul, omdat de eindbestemming fout is? Een kandidaat die de eerste instructie fout opvolgt, zal veel minder punten scoren dan een kandidaat die pas bij de laatste instructie in de fout gaat. Is dit de bedoeling? Het antwoord op die laatste vraag kan 'ja' zijn. Misschien vindt u het juist belangrijk dat de kandidaat zo dicht mogelijk bij het eindpunt uitkomt en verdient de tweede kandidaat dus effectief meer punten dan de eerste.

De preconditie: doet hij wat hij moet doen?

Om te voorkomen dat een kandidaat louter geslaagd is op basis van zijn score voor vormelijke aspecten, terwijl de inhoud niet aansluit bij de opdracht, kunt u een zogenaamde **preconditie** voorzien. Het kan gebeuren dat een kandidaat zich voorbereidt op een argumentatietaak, door een vormelijk zeer goed uitgebouwde argumentatie uit het hoofd te leren. Een vergaand maar duidelijk voorbeeld: een kandidaat krijgt de opdracht om te argumenteren waarom het voor bejaarden beter is om bij familie in te trekken dan om in een verzorgingstehuis te gaan wonen, maar hij houdt een pleidooi voor het openbaar vervoer en de fiets en tegen de auto (bvb. omdat hij dat in de les geoefend heeft). De beoordelaar merkt dat de inhoud niet aansluit bij de opdracht. Anderzijds kan het ook gebeuren dat een kandidaat een antwoord geeft dat helemaal niet aansluit bij de opdracht. Het is dan een zeer frustrerende bezigheid om de hele prestatie te beoordelen. Door een preconditie te voorzien, bv. "de kandidaat geeft geen argumentatie of de argumentatie sluit niet aan bij de opgegeven stelling = score 0 voor de hele taak", is dit probleem van de baan.

Als u zich een goed beeld heeft gevormd van welke criteria van belang zijn, komt de volgende stap: hoe brengen we die criteria precies in kaart?

HOE?

Hoe kunnen we het best de concrete informatie verkrijgen? Hoe komen we te weten wat we precies willen weten?

Wacht ten eerste niet met het beoordelingsmodel tot de (toets)taak af is, maar **ontwikkel taak en beoordelingsmodel tegelijk**. Het is het perfecte moment om bij elke opgave te bepalen wat een goed antwoord is, hoeveel punten dat oplevert en wat als een fout antwoord wordt beschouwd. Op die manier kan de toets(taak) ook onafhankelijk van de constructeur functioneren. Niet alleen u maar ook anderen (bvb. collega's) kunnen de toets dan volgens dezelfde criteria objectief en consequent scoren. Het beoordelingsmodel geeft heel precies aan wat goed en wat fout is en laat zo weinig mogelijk ruimte voor interpretatie en subjectiviteit. Om heel goed de vinger op de wonde te kunnen leggen, zijn meer nauwkeurige omschrijvingen nodig van de geselecteerde criteria. Het ERK, of de taalvaardigheidseisen of doelstellingen, zijn niet alleen een goede basis voor de selectie van beoordelingscriteria, maar ook voor de concrete formulering ervan. In het voorbeeld van de argumentatietaak, namen we in het analytische model 'spelling' op als een relevant criterium. Om precies te omschrijven wat we bedoelen met 'spelling' voor deze taak op C1-niveau, kunnen we de omschrijving uit het ERK (blz. 110) gebruiken:

	ORTHOGRAFISCHE BEHEERSING
C2	*Het geschrevene is orthografisch foutloos.*
C1	*Lay-out, alinea-indeling en leestekengebruik zijn consistent en bevorderen de leesbaarheid. De spelling is correct, afgezien van een enkele verschrijving.*
B2	*Kan heldere, begrijpelijke, doorlopende tekst produceren die voldoet aan standaardconventies voor lay-out en alinea-indeling. Spelling en leestekengebruik zijn redelijk correct maar kunnen invloeden van moedertaal verraden.*
B1	*Kan heldere doorlopende tekst produceren die over het algemeen helemaal te begrijpen is. Spelling, leesteken-gebruik en lay-out zijn correct genoeg om het grootste deel van de tijd te kunnen worden gevolgd.*
A2	*Kan korte zinnen over alledaagse onderwerpen overschrijven, bijvoorbeeld een routebeschrijving. Kan fonetisch redelijk correct (maar niet noodzakelijkerwijs helemaal in de standaardspelling) korte woorden opschrijven uit zijn of haar gesproken woordenschat.*
A1	*Kan vertrouwde woorden en korte frasen overschrijven, bijvoorbeeld eenvoudige aanwijzingen of instructies, namen van alledaagse dingen, namen van winkels en regelmatig gebruikte standaardzinnetjes. Kan zijn of haar adres, nationaliteit, en andere persoonlijke gegevens spellen.*

Deze verwoording kunt u dan herschrijven in functie van wat u zelf relevant acht voor deze specifieke toetssituatie voor deze specifieke kandidaten.

Wanneer je een beoordelingsmodel concreet uitwerkt, wordt ook aandacht besteed aan de verdeling van de onderdelen inhoud en vorm en aan de verdere onderverdeling in het toekennen van een waarde per criterium, bv. dichotoom (0 of 1) of polytoom (bv. een getal van 1 tot 4), of een waardering in woorden. Om deze beslissing te nemen, speelt ook mee welk 'gevoel' overheerst bij het scoren van een prestatie: misschien volstaat het om 0 of 1 te scoren, misschien ervaart u de nood aan een meer 'gedifferentieerd' oordeel en wenst u daarom de keuze tussen een waarde van 1 tot 4. Een praktische tip hierbij is dat het goed is om een even aantal keuzemogelijkheden te voorzien. Een beoordelaar die kan kiezen tussen scores 1, 2 en 3, heeft de neiging zeer vaak voor de 2 te kiezen.

Daarnaast komt hier het element '**weging**' aan de orde. Stel dat u de nadruk wilt leggen op inhoudelijke aspecten en niet op vormelijke. In een globaal beoordelingsmodel, kunt u dat doen door in de omschrijving van de verwachting de inhoudelijke elementen te benadrukken. Bij een analytisch model is het mogelijk om bv. zeven of acht inhoudelijke items te scoren en twee of drie vormelijke. Het is ook mogelijk dat u een analytisch model hebt opgesteld met even veel inhoudelijke als vormelijke criteria. Dan kunt u beslissen om de score op de inhoudelijke criteria zwaarder te laten doorwegen door die score bv. met een factor twee te vermenigvuldigen. Idealiter wordt een wegingsfactor bepaald op basis van statistische analyses, maar dit leidt ons te ver voor deze toetstip.

Om een beoordelingsmodel op te stellen dat zo weinig mogelijk ruimte tot interpretatie en twijfel laat, kunnen de volgende tips helpen:

- Ook na de eerste afname van de toets is het erg belangrijk om je beoordelingsmodel in de gaten te houden. Vaak krijg je dan nog beter zicht op wat de ideale prestatie op een taak is én op antwoorden die mogelijk leiden tot twijfel bij de beoordelaars: is het een goed of een fout antwoord? Bijvoorbeeld: u vraagt studenten schriftelijk hun mening te formuleren over een bepaald onderwerp. Sommige studenten beschrijven de situatie veeleer objectief zonder expliciet hun mening te formuleren. Of een student geeft in zijn inleiding aan 'pro' te zijn en geeft vervolgens een argument 'contra'. Rekent u hun antwoord goed of fout? Zorg ervoor dat uw beoordelingsmodel in deze gevallen strikt aangeeft wat u als een goed en wat u als een fout antwoord moet beschouwen. Echte kandidaatsprestaties kunnen met andere woorden helpen om je beoordelingsmodel uit te breiden en scherp te stellen.
- Geef, eventueel in een ander document, per vraag en opdracht voorbeelden van goede en foute antwoorden. Een dergelijk 'beslisboekje' kan een zeer nuttig naslaginstrument zijn om soortgelijke beoordelingsmodellen te ontwikkelen.
- Om na te gaan of uw analytische model goed overeenkomt met wat u van een kandidaatsprestatie verwacht, kunt u aan de prestatie een globaal oordeel toekennen. Ligt uw globale oordeel doorgaans hoger dan het analytische, dan zijn bepaalde items in het analytische model misschien te moeilijk. Omgekeerd zijn punten misschien te gemakkelijk te verdienen voor bepaalde analytische items, als het analytische oordeel consequent hoger ligt dan het globale. Hou echter ook steeds uw mildheid en/of die van de beoordelaar(s) in de gaten. Als uw globaal oordeel bvb. steevast hoger ligt dan het analytische, kan het zijn dat u als beoordelaar te mild bent.

EN DAN?

Dan ligt er een eerste versie van uw beoordelingsmodel. Mogelijk is dat een zeer goede versie, maar de ervaring leert dat het veel waarschijnlijker is dat er nog aan gesleuteld moet worden op basis van ervaringen bij de eerste afname. De laatste tip hierboven kan namelijk ontluisterende resultaten opleveren en grondige aanpassingen vragen.

Ten slotte is het belangrijk om na te gaan:

- …of er voldoende samenhang is tussen toetsopzet en beoordeling (bv. in de instructies voor de kandidate; weten zij waarop zij beoordeeld worden?)

- …hoe je de resultaten kunt interpreteren en hoe je daarover het best communiceert met kandidaten.
- …welke informatie de studenten krijgen over hun score, hoe die tot stand is gekomen (de beoordeling) en wat ze daaruit kunnen leren voor hun vorderingen (diagnostische info).

In het andere geval moet nog beslist worden waar de cesuur zal liggen: wie zakt en wie slaagt op basis van het model? Waar ligt de grens? Of waar liggen de grenzen, als de kandidaten in meerdere categorieën ingedeeld moeten worden, zoals bijvoorbeeld bij een instaptoets? Bij grote examens met veel kandidaten wordt de beslissing omtrent de cesuur genomen op basis van uitgebreide statistische analyses en expertoordelen. Voor de cesuurbepaling van een toets voor een kandidatengroep die u goed kent, kunt u echter aan de slag op basis van uw ervaring. U houdt dan de zogenaamde 'minimaal acceptabele kandidaat' voor ogen en beslist welke items die virtuele kandidaat minimaal goed moet maken om geslaagd te zijn. Zo bepaalt u het strikte minimum van het te behalen niveau op de toets, dat vertaald wordt in de cesuur. Die kan bestaan uit één eindscore (bv. '7/10'), of uit een aantal voorwaarden om te slagen, bv. 'de kandidaat moet slagen voor elke taak afzonderlijk om geslaagd te zijn voor het geheel van de toets' of 'de kandidaat moet minimaal items 2, 4 en 5 goed hebben om te kunnen slagen'.[i]

De definitieve versie

Om het beoordelingsmodel definitief vast te stellen vindt er idealiter een pilootafname plaats. Aan de hand van de ervaringen die opgedaan zijn tijdens de pilootafname kan het model aangepast worden voor het 'echt' wordt ingezet.

Als er bij de uiteindelijke beoordeling meerdere beoordelaars zijn, is een beoordelaarstraining strikt noodzakelijk. De training is in de meeste gevallen een soort 'oefenbeurt' waarbij alle deelnemers pilootprestaties beoordelen en vervolgens hun score vergelijken en bespreken, met als doel op dezelfde golflengte te komen. Dat kan gebeuren op basis van een eerder afgenomen parallelle taak, ofwel op basis van een pilootafname bij een vergelijkbare groep studenten, bv. van een bevriende instelling.

Om bij de 'echte' beoordeling na te gaan of iedereen daadwerkelijk op dezelfde golflengte zit, kan een dubbele beoordeling plaatsvinden (twee beoordelaars beoordelen dezelfde prestaties), of indien dat qua tijd niet kan: een individuele beoordeling met overleg over alle twijfelgevallen.

Het is interessant om zowel van de training als van de bevindingen bij de 'echte' beoordeling in een logboek bij te houden welke punten discussie opriepen, welke prestaties als twijfelgeval aangemerkt werden enz.

Een gouden tip doorheen het hele ontwikkelproces, is "doe dit niet alleen". Een beoordelingsmodel ontwikkelen is zeer arbeidsintensief en tijdsrovend. De kwaliteit van het model gaat er zeker op vooruit door met verschillende mensen samen te werken en het model te herwerken tot iedereen tevreden is. Dat komt de betrouwbaarheid van uw instrument ten goede en is dus ook goed voor uw kandidaten.

STAPPENPLAN

1. WAAROM

Keuze soort model op basis van doel (plaatsingstoets of diagnostische toets) en op basis van grootte van de te beoordelen groep.

2. WAT

Selectie van de criteria die opgenomen worden in het model. Inhoudelijke, vormelijke en andere aspecten selecteren.

3. HOE

Precieze omschrijving van de geselecteerde criteria en kiezen van een waardesysteem (dichotoom, polytoom, cijfers, letters?). Eventueel weging voorzien.

4. Inzet in de praktijk

Indien nodig: cesuurbepaling. Optimaliseren op basis van proefafname. Beoordelaarstraining en controle bij uiteindelijk gebruik.

[i] [i] Meer uitleg bij de cesuurbepaling, vindt u in de betreffende toetstip. Die kunt u vinden op www.CNaVT.org en doorklikken in het archief van de toetstips (2006-2007 toetstip 10).