

HANDREIKINGEN VOOR IMPLEMENTATIE VAN DIGITAAL TOETSEN MET BEELDEN

Anouk van der Gijp
Cécile Ravesloot
Marieke van der Schaaf
Kim de Crom
Corinne Tipker
Christian Mol
Joost van den Berg
Irene van der Schaaf
Pieter-Jan van der Schoot
Dik Rutgers
Mario Maas
Jan van Schaik
Koen Vincken



Inhoud

Inleiding en begrippenlijst	5
H1. ICT-infrastructuur	8
H2. Ontwerpen van een digitale radiologietoets	16
H3. Feedback bij toetsing van radiologische beeldinterpretatie	24
H4. Praktische aandachtspunten implementatie digitale radiologietoetsing	28
H5a. Evalueren van een digitale radiologietoets	33
H5b. Verduurzaming van een digitaliseringsproject	38
Referenties	42

Inleiding

In de medische specialistenopleidingen zijn beeldinterpretatievaardigheden van groot belang. Binnen de radiologie kunnen beelden tegenwoordig zowel tweedimensionaal (2D) als meerdimensionaal (multiplanar reformatted beelden, afgekort MPR-beelden) worden bekeken en beoordeeld. Echter, de toetsing van beeldinterpretatievaardigheden liep tot op heden achter. De toetsing vond plaats op een arbeidsintensieve, schriftelijke, en daarmee weinig authentieke, wijze. Uit het SURF-project 'Toetsing in beeld' is gebleken dat de kwaliteit van toetsing van beeldinterpretatievaardigheden verbeterd kan worden door middel van digitale en authentiekere toetsing met 2D- en MPR-beelden. In het project 'Digitaal toetsen met beelden, een extra dimensie' is digitaal toetsen met 2D- en MPR-beelden geïmplementeerd in de medische specialistenopleiding radiologie op landelijke schaal. Hiervoor is het digitale toetsprogramma VQuest gebruikt dat speciaal ontwikkeld is om beeldinterpretatievaardigen te toetsen met zowel 2D- als MPR-beelden. Het toetsprogramma maakt het mogelijk om te toetsen met behulp van volume datasets. Kandidaten kunnen door de beelden heen scrollen in meerdere richtingen en er zijn verschillende andere beeldbewerkingsopties mogelijk, bijvoorbeeld zoomen en contrastinstellingen aanpassen. De implementatie van digitaal toetsen met beelden (o.a. MPR-beelden) vergt een specifieke organisatie en ICT-infrastructuur. In dit handboek wordt ingegaan op de volgende aspecten voor het ontwikkelen en implementeren van digitale toetsen met beelden: ICT randvoorwaarden (H1), ontwerpen van een digitale radiologietoets (H2), feedback voor toetsing van radiologische beeldinterpretatie (H3), aandachtspunten bij de implementatie digitale radiologietoetsing (H4), evalueren van een digitale radiologietoets en verduurzaming van de digitalisering (H5). Dit wordt onderbouwd en geïllustreerd aan de hand van voorbeelden uit het project, gericht op de integratie van digitaal toetsen van 2D- en MPR-beeldinterpretatie in de landelijke voortgangstoets radiologie.

Begrippenlijst

2D-beeld: Een 2D-beeld is één geselecteerde doorsnede uit een MPR-beeld. Hierbij is het niet mogelijk om het beeld in andere richtingen te bekijken. Een 2D-beeld kan ook een projectiebeeld van een 3D structuur zijn, zoals een röntgenfoto.

2D- en MPR-beeldvraag: Dit is een toetsvraag aan de hand van een 2D- respectievelijk MPR-beeld.

Aanwijsvraag: Toetsvraag waarbij de kandidaat gevraagd wordt een structuur of afwijking in een beeld te lokaliseren en te markeren.

Alternatievenlijst: Lange lijst met antwoorden voor longmenuvraag.

Arts-assistenten / aios: Artsen in opleiding tot specialist

Assemblage: selectie van toetsvragen en toetssamenstelling

Beeldinterpretatie: Het beoordelen van beelden leidend tot een conclusie over de betekenis van het beeld.

Construct: Hetgene wat men wil toetsen, dat wil zeggen de geïntegreerde kennis, begrip en vaardigheden die men zou moeten beheersen voor het kunnen uitoefenen van een bepaalde taak, bijvoorbeeld beeldinterpretatie.

Constructvaliditeit: Een indicatie of de test de kennis, het begrip en/of de vaardigheden meet die men wil meten. Zie voor een uitgebreide uitleg hoofdstuk 5a2.

Contrastvenster/-instellingen: Instellingen voor window-level en window-width. Deze instellingen geven weer welke contrasten door de range van grijswaarden worden weergegeven.

Coupe: één doorsnede uit een stack serie (zie *stack*).

CPU: Central Processing Unit. Dit is de centrale verwerkingseenheid van de computer, waar het rekenwerk plaatsvindt.

Diagnosevraag: Toetsvraag waarbij de student gevraagd wordt aan de hand van een radiologisch beeld de meest waarschijnlijke diagnose te stellen.

Differentiaal diagnose: Lijst van mogelijke *diagnoses* in volgorde van waarschijnlijkheid (opgesteld op grond van bijvoorbeeld een bepaalde klachtenpresentatie of afwijkend radiologisch beeld).

Formatieve toetsing: Toetsing die bedoeld is voor het verzamelen van leerresultaten voor het geven van feedback, met als doel het leerproces van kandidaten te stimuleren en te sturen. Doorgaans heeft het toetsresultaat geen consequenties voor de studievoortgang van de kandidaten.

HORA-opleidingsplan: Opleidingsplan radiologie, opgesteld door de commissie Herziening Opleiding Radiologie in 2010.

Inhoudsvaliditeit: Een indicatie of de toets representatief is voor hetgeen getoetst moet worden en of de toets het onderwerp voldoende dekt. Ofwel, is de toets een juiste en gelijkmatige afspiegeling van de te meten kennis en vaardigheden zoals deze gedefinieerd zijn. Zie voor een uitgebreide uitleg hoofdstuk 5a2.

Longmenuvraag: Toetsvraag waarbij de kandidaat uit een zeer lange lijst met antwoorden (*alternatievenlijst*) het juiste antwoord kan kiezen.

Modaliteiten: Radiologische beeldvormende technieken. Voorbeelden van verschillende radiologische beeldvormende technieken zijn: echogram (beeld vervaardigd met echografie), röntgenfoto (2D-beeld vervaardigd met röntgenstraling), CT-scan (dwarsdoorsnede beeldvorming met behulp van röntgenstraling) en MRI-scan (dwarsdoorsnede beeldvorming met behulp van magnetische resonantie).

Multiplanar reformatted (MPR) beeld: Dit beeld betreft een volume dataset, bestaande uit vele opeenvolgende dwarsdoorsneden van het menselijk lichaam. De beoordelaar kan door de dwarsdoorsneden heen scrollen in diverse richtingen. Dit in tegenstelling tot een 2D-beeld waarbij één doorsnede geselecteerd is uit de 3D volume dataset.

Opleider: De opleider is de radioloog die verantwoordelijk is voor de opleiding van de arts-assistenten in zijn opleidingskliniek.

Opleidingskliniek: Een ziekenhuis waarin artsen worden opgeleid tot specialist. Een specialistenopleiding is verdeeld over meerdere *opleidingsklinieken*.

Overlay: Gesegmenteerde structuur in het beeld door middel van het 'inkleuren' van de structuur in 2D of in een volume dataset. Hiermee kunnen bijvoorbeeld aanwijsvragen worden nagekeken.

Radiologisch beeld/onderzoek: Beeld dat met behulp van een radiologische techniek wordt gemaakt van (een deel van) het menselijk lichaam. Er zijn verschillende radiologische onderzoeken, dit worden ook wel *modaliteiten* genoemd (zie voorbeelden onder het kopje *modaliteiten* in deze begrippenlijst).

Radiologische beeldinterpretatie: Het proces van het interpreteren van radiologische beelden hetgeen leidt tot een conclusie, bijvoorbeeld in de vorm van een (differentiaal) diagnose.

Stack: serie doorsnede beelden van een driedimensionaal object (lichaamsdeel), waar men doorheen kan scrollen.

Studentmodel: Gedetailleerde beschrijving van de vaardigheid die getoetst moet worden in hoofd- en subonderdelen.

Summatieve toetsing: Toetsing waarvan het toetsresultaat consequenties heeft voor de studievoortgang van de studenten.

Taakmodel: Beschrijving van de kenmerken die toetsvragen moeten hebben om aan te kunnen geven in hoeverre kandidaten de vaardigheden (die men beoogt te toetsen) beheersen.

Taxonomie van Bloom: De taxonomie van Bloom is een classificatiesysteem dat gebruikt wordt voor het definiëren en onderscheiden van verschillende beheersingsniveau's: kennis, begrip, toepassen, analyseren, evalueren en synthetiseren.

Viewport: Een gedeelte van de userinterface dat gebruikt kan worden voor het weergeven van afbeeldingen. Indien er meerdere viewports zijn is de userinterface dat gereserveerd is voor het weergeven van beelden verdeeld, zodat meerdere afbeeldingen tegelijk kunnen worden weergegeven.

Voortgangstoets (VGT): Toetsvorm die de voortgang van kandidaten vaststelt door het herhaaldelijk afnemen van een toets (van ongeveer gelijk niveau en inhoud) tijdens een opleidingstraject. Deze toetsvorm wordt meestal gebruikt voor formatieve doeleinden.

VQuest: Digitaal toetsprogramma speciaal ontwikkeld voor het toetsen met beelden. Het biedt de mogelijkheid tot diverse beeldmanipulaties tijdens het maken van de toets, zoals bv. het wisselen van kijkrichting (bij volume datasets) en contrastinstellingen.

Hoofdstuk 1. ICT-infrastructuur

1.1 Inleiding

Digitale toetsing met MPR-beelden wordt verondersteld efficiënter en authentieker te zijn dan traditionele schriftelijke toetsing en daarmee bij te dragen aan een verbetering van de toetskwaliteit. Dit gaat echter alleen op als de toets zonder verstoringen kan worden afgenomen. Dit betreft niet alleen de betrouwbaarheid van de verkregen scores (fraudemogelijkheden moeten bijvoorbeeld worden voorkomen), maar ook de betrouwbaarheid van het onderliggende systeem en de techniek. Dit vereist dat alle randvoorwaarden, die nodig zijn voor een betrouwbaar toetssysteem, worden vastgelegd en getest met behulp van stresstests. Tevens moeten de praktische bruikbaarheid, efficiëntie, snelheid en veiligheid van het systeem getest zijn.

Voor het toetsen met digitale 2D- en MPR-beelden is in dit project gebruik gemaakt van het programma VQuest, een digitaal toetsprogramma dat is ontworpen voor toetsing van (radiologische) beeldvragen. In VQuest is het mogelijk om te toetsen aan de hand van volume datasets. Kandidaten kunnen door de datasets (bijvoorbeeld een CT-scan) heen scrollen in meerdere richtingen. Daarnaast zijn er verschillende andere datamanipulaties mogelijk, die ook in de praktijk worden gebruikt, zoals in- en uitzoomen en contrastinstellingen wijzigen. In dit hoofdstuk worden de randvoorwaarden beschreven die zijn onderzocht en vastgesteld in het project 'Digitaal toetsen met beelden, een extra dimensie' voor het succesvol afnemen van digitale toetsen met 2D- en MPR-beelden op grote schaal. In het bijzonder wordt ingegaan op welke eisen dit stelt aan de ICT-infrastructuur, ten aanzien van de toetszaal en werkplekken, het netwerk- en systeembeheer, het toetsprogramma, het uitwisselen van toetscontent en het te gebruiken nakijkmodel. In paragraaf 1.4 zullen de specifieke eisen aan het toetsprogramma VQuest worden toegelicht.

1.2 Toetszaal

Voordat besloten wordt of een toetszaal geschikt is voor afname met VQuest, moet aan een aantal voorwaarden zijn voldaan.

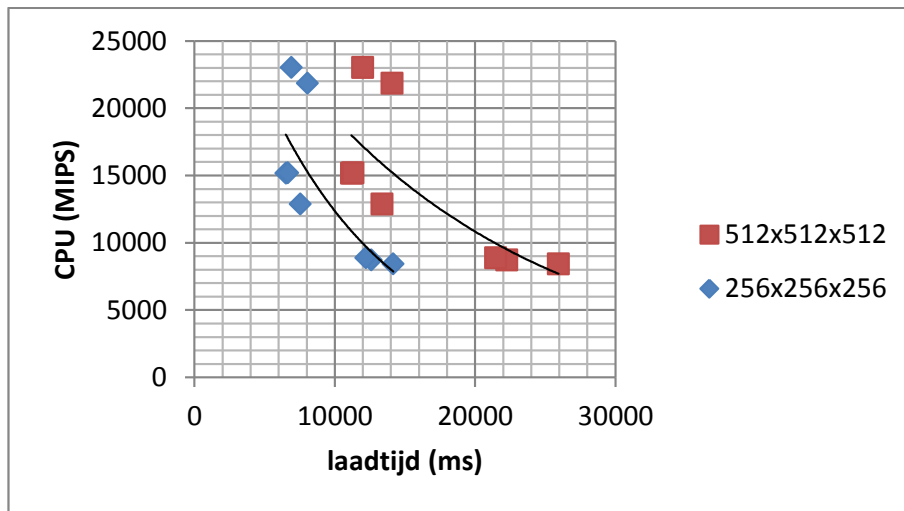
- *Minimaal aantal computers*

Voor digitale toetsafname is het van belang rekening te houden met mogelijke defecten van computers. Uit de diverse toetsen die al zijn afgenomen met VQuest, is gebleken dat een marge van ongeveer 2 à 3% extra computers nodig is, oftewel 10 reservecomputers op 400 kandidaten. Uitval van computers heeft verschillende redenen. Zo kan enerzijds de computer zelf een defect vertonen (bijvoorbeeld uitval van de monitor, hardware of muis), anderzijds kan een werkplek ongeschikt blijken te zijn door bijvoorbeeld storende lichtinval of geluidsoverlast.

- *Minimale hardware eisen*

VQuest draait momenteel (2014) op alle systemen met Windows XP t/m Windows 8. De (gepercipieerde) snelheid van VQuest wordt vooral bepaald door de laadtijd van de (grotere) beeld-datasets. Hierbij is de vuistregel dat het laden van één dataset niet langer mag duren dan 10 seconden. De laadtijd daalt ongeveer exponentieel met de snelheid van de CPU (zie Figuur 1a). Een CPU met een kracht van 10.000 MIPS¹ is werkbaar gebleken in de praktijk, maar kan bij grotere datasets (of bij gebruik van meerdere viewports in één vraag) tot iets te lange laadtijden leiden. Aangeraden wordt om een CPU te gebruiken met minimaal 15.000 MIPS. Het interne geheugen moet minimaal 4 GB bedragen. Er moet een grafische kaart aanwezig zijn van voldoende kwaliteit, bijvoorbeeld NVIDIA met 1 GB geheugen.

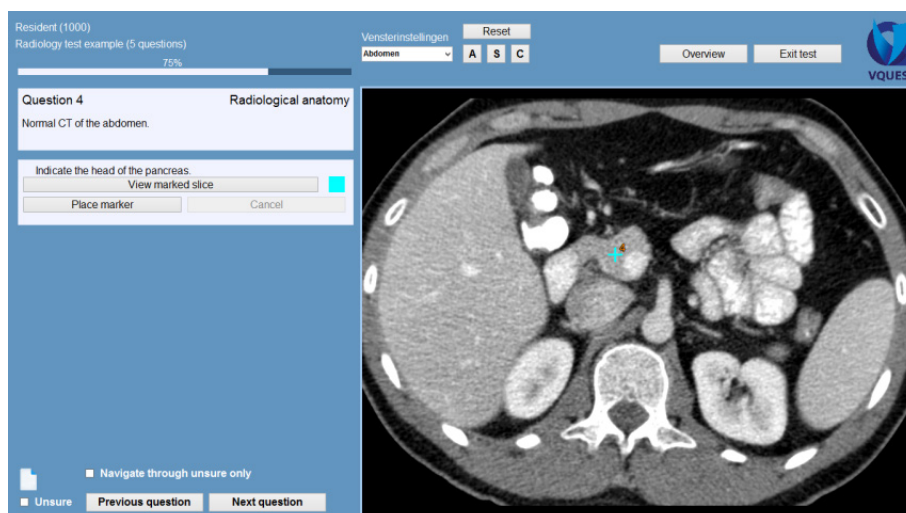
¹Het aantal MIPS van de CPU kan bijvoorbeeld bepaald worden met een benchmark programma als Dacris (www.dacris.com).



Figuur 1a. Verband tussen de snelheid van een CPU en de laadtijd in ms voor twee datasets van verschillende grootte (256^3 en 512^3 voxels)

- **Minimale schermresolutie**

Bij het bepalen van de minimale schermresolutie spelen twee zaken een rol: (1) fysieke afmeting van de gebruikte monitor; (2) gebruikte resolutie. In de praktijk blijkt dat een fysieke beeldscherm afmeting van minimaal 15.6" wenselijk is, om te voorkomen dat kandidaten steeds in moeten zoomen om details zichtbaar te maken. Veel inzoomen gaat ten koste van de toetstijd en komt niet overeen met de dagelijkse praktijk van een praktiserend radioloog (en wordt hierdoor als belastend ervaren door aios). De gebruikte resolutie is van invloed op de lay-out van de userinterface van de VQuest applicatie (Figuur 1b). De interface kan grofweg worden opgedeeld in drie delen: de bovenste balk (met algemene informatie over de toets en visualisatie mogelijkheden), het linkerframe (met alle subvragen en navigatiemogelijkheden), en het rechterframe (beelddata). VQuest is zo ontworpen dat de bovenste balk altijd een vaste hoogte en het linkerframe altijd een vaste breedte heeft. Dat is gedaan om het lezen van de vraag zo duidelijk en consistent mogelijk te houden. Dit heeft als consequentie dat als de resolutie van het scherm lager wordt in horizontale richting (dus in de breedte van het scherm), dit ten koste gaat van de breedte van het plaatje. Een resolutie van 1366x768 pixels wordt als absoluut minimum beschouwd. Deze resolutie is diverse keren gebruikt bij toetsen gedurende de looptijd van het project, maar wordt door de gebruikers (zowel docenten als kandidaten) als matig beschouwd. Aanbevolen resoluties zijn WSXGA+ (1680x1050), of UXGA (1600x1200), of WUXGA (1920x1200).



Figuur 1b. Screenshot van een aanwijsvraag in VQuest.

- *Omgevingsfactoren tijdens toetsen (geluid, lichtinval etc.)*

Elke toetszaal moet voldoen aan een aantal basiseisen. Het omgevingsgeluid moet acceptabel zijn, de temperatuur moet acceptabel zijn en de zaal moet enigszins verduisterd kunnen worden. Met name bij radiologie is het een feit dat de beelddata het beste bekeken kan worden in een relatief donkere omgeving. Het is voorgekomen dat kandidaten wegens storend lichtinval op de monitor tijdens een toets verplaatst moesten worden.

- *Zaalindeling (i.v.m. fraude)*

Ter voorkoming van fraude kunnen verschillende maatregelen worden getroffen. Het uitgangspunt hierbij is dat kandidaten weinig tot geen zicht hebben op de beeldschermen van collega-kandidaten, waardoor het risico op afkijken wordt geminimaliseerd. Dit kan bijvoorbeeld door de kandidaten aan aparte tafeltjes te laten plaats nemen, zoals in de digiTenT (zie Foto 1). Aanvullend zouden er schotten kunnen worden geplaatst tussen werkplekken of kan er gebruik worden gemaakt van een speciale beeldschermfolie. Een nadeel van het gebruik van schotten is dat de werkplekruimte van de kandidaat afneemt. Bij het gebruik van speciale beeldschermfolie, waardoor je het beeldscherm alleen kunt bekijken als je er min of meer recht voor zit, verkleuren de randen van het beeldscherm daarentegen enigszins. Hierdoor moet men soms gaan verzitten om toch het gehele beeld goed te kunnen beoordelen. VQuest heeft de optie om alle vragen in willekeurige volgorde aan te bieden. Dat verkleint de kans aanzienlijk dat twee naburige kandidaten met dezelfde vraag beginnen of bezig zijn. Echter, aangezien de beelden ervoor kunnen zorgen dat een vraag makkelijk herkenbaar is door medekandidaten, bevelen we aan om (bovenstaande) aanvullende maatregelen te treffen ter voorkoming van fraude.



Foto 1. Impressie van een toetsafname in de digiTenT

- **Softwareomgeving**

VQuest draait lokaal op elke computer die gebruikt wordt voor toetsafname. Technische support is nodig om de toets over de verschillende werkplekken te distribueren. Tijdens de toets zelf is het wegschrijven van de antwoorden en logfiles naar de server feitelijk het enige netwerkverkeer. De antwoorden worden op twee verschillende plekken opgeslagen: lokaal en op een server. De lokale map dient als back-up. De antwoorden op de server kunnen tijdens de toets bekeken worden door de technisch beheerder, zodat duidelijk is dat het systeem werkt en dat de antwoorden correct worden opgeslagen. Hiervoor is ook een monitor-applicatie beschikbaar, waarmee inzichtelijk wordt gemaakt hoeveel antwoorden er binnen zijn gekomen, en dus ook wat de status is van alle kandidaten bij het maken van de toets.

- **Aantal locaties / aantal computerzalen / aantal computers per locatie**

Voor het gebruik van VQuest heeft één zaal op één locatie de voorkeur. Het gebruik van meerdere zalen en/of meerdere locaties heeft als gevolg dat VQuest vaker getest moet worden en dat de toetsdistributie en beveiliging op alle locaties en in alle zalen georganiseerd en getest moeten worden. Tijdens dit SURF-project is de eerste VGT met VQuest afgenomen op twee verschillende locaties: AMC (Amsterdam) en UMC Utrecht. Per locatie waren verschillende toetszalen nodig (vijf zalen in het UMCU en acht zalen in het AMC) om aan de totale vraag te voldoen. De tweede en derde VGT (respectievelijk november 2013 en april 2014) zijn in één grote zaal afgenomen: de digiTent op het terrein van de VU, Amsterdam. In hoofdstuk 4 worden de ervaringen van deze toetsafnames beschreven.

1.3 Netwerk- en systeembeheer

VQuest is een *stand-alone* applicatie die niet onder Windows geïnstalleerd hoeft te worden, en slechts gekopieerd hoeft te worden. Dat betekent dat het distribueren van zowel de software (VQuest) als de toets (content) voldoende is om op elke computer gedraaid te kunnen worden. Hiertoe moet een voorziening aanwezig zijn om deze distributie soepel te laten verlopen. Aangezien een toetszaal normaliter gehuurd moet worden, verdient het de aanbeveling de verspreiding van de toets binnen 1 uur af te ronden. In de praktijk is gebleken dat er twee systemen bruikbaar zijn:

1. Distributie vanuit een server door op elke computer in te loggen met een VQuest-account, waarna alle benodigde files automatisch gekopieerd worden naar een lokale schijf op de computer.
2. Distributie via een system management oplossing, waarbij zonder inloggen de toets al op elke computer gezet kan worden (op afstand).

Er wordt bij voorkeur gebruik gemaakt van een bedraad netwerk. De reden is tweeledig. Ten eerste zijn draadloze netwerken minder goed te beveiligen, en ten tweede is de benodigde bandbreedte beperkt, terwijl er mogelijk enkele Gigabytes aan data over het netwerk gekopieerd moeten worden naar meerdere werkplekken. Ter illustratie: tijdens de eerste toetsafname bij arts-assistenten met VQuest (april 2013) is noodzakelijkerwijs gebruik gemaakt van extra laptops, omdat er te weinig vaste werkplekken waren. Bij het opstarten en kopiëren van VQuest bleek dat de router de netwerkbelasting niet aan kon. Er is toen besloten de toets ter plekke met USB sticks te verspreiden. Hoewel werkbaar, is dit geen optimale oplossing, omdat het veel tijd en mankracht vergt.

Beveiliging

Los van de distributie moet de toetsomgeving beveiligd zijn tegen gebruik van andere applicaties (zoals Internet Explorer). Beveiliging kan onder andere met kiosksoftware (afschermingssoftware) of via *group policies* van Windows. Hiermee is precies te regelen welke applicaties gestart mogen worden en zijn bepaalde toetscombinaties af te schermen die toegang geven tot onderdelen van Windows die ongewenst zijn tijdens toetsafname. Het netwerk zou eventueel geconfigureerd kunnen worden, zodat communicatie met het Internet in zijn geheel niet mogelijk is.

Feitelijk is het enige netwerkverkeer het wegschrijven van de antwoorden en logfiles naar de server tijdens de toets. Tijdens het testen van VQuest op locatie moet er speciaal op gelet worden dat alle antwoorden en logfiles op de juiste plek worden weggeschreven in de beveiligde omgeving. Het heeft de voorkeur de antwoorden op twee plaatsen op te slaan: op de server en lokaal op de computer. De lokaal weggeschreven antwoorden dienen als back-up. In sommige toetszalen is het mogelijk om elke computer afzonderlijk “over te nemen”, of alleen mee te kijken. Ook kan soms een mededeling centraal verspreid worden vanuit een serverruimte. Hiervoor is netwerkpermissie een vereiste.

1.4 Acceptatietesten toetsprogramma VQuest

Het toetsprogramma VQuest is continu in ontwikkeling. Voor een toets moet een stabiele versie getest worden en daarna niet meer aangepast worden. Om problemen te voorkomen is de aanbeveling om uiterlijk 2 maanden voor de toets de VQuest versie te bevriezen. Alleen bij serieuze lacunes kan dan nog besloten worden om de software aan te passen.

Na elke toets kan de evaluatie leiden tot een aanpassing van VQuest. In het verleden is bewezen dat door de aanpassingen van VQuest de toets steeds beter werd. Bijvoorbeeld de snelheid van het laden van de MPR datasets werd steeds positiever beoordeeld in de gebruikersevaluaties.

Voorkomen moet worden dat zich een probleem tijdens de toetsafname zelf voordoet. Dit risico kan beperkt worden door uitgebreid te testen. Daarbij is de aanbeveling om op 4 verschillende niveaus te testen:

1. *Op een losse computer, antwoorden lokaal wegschrijven.*
Hierbij wordt vooral getest of de toetssoftware goed werkt. Dit betreft het opstarten van VQuest, het inlezen van de toets, het wisselen tussen de vragen, het manipuleren van de beelddata, en het afsluiten van VQuest.
2. *Op een losse computer, antwoorden naar een server via het netwerk.*
Hierbij wordt vooral gekeken of de antwoorden goed worden weggeschreven naar de server. Bij het herstarten van de toets onder dezelfde login, worden de eerder gegeven antwoorden van de server ingelezen. Dit moet vlekkeloos werken bij inloggen op dezelfde en op een andere computer.
3. *In een netwerk van computers, antwoorden lokaal.*
In deze fase moet vooral gekeken worden of de netwerkbeveiliging goed werkt. Meerdere computers moeten dan tegelijkertijd opgestart worden met VQuest en tegelijkertijd getest.
4. *In een netwerk van computers, antwoorden naar een server.*
Dit is de finale test. Alle opties van VQuest (antwoorden via server, back-up lokaal, opnieuw inloggen op andere computer, afsluiten VQuest) moeten dan werken.

Bij elk van bovenstaande testen wordt gekeken naar de volgende punten:

- a. *Opstarttijd Toets*
De opstarttijd mag niet onnodig lang zijn en is mede afhankelijk van de beschikbare voorbereidings-tijd op de locatie waar de toets wordt afgenomen. Het distribueren en opstarten van de toets moet bij voorkeur in maximaal 1 uur afgerond zijn.
- b. *Laadtijd dataset*
De maximale laadtijd voor een grote dataset (het beeldmateriaal voor één vraag) is ongeveer 10 seconden (zie ook 1.1). Omdat het voor kan komen dat een computer of laptop deze limiet niet haalt, is in VQuest een laadbalk opgenomen die het verstreken percentage van de laadtijd aangeeft. Daarmee wordt een eventuele langere laadtijd eerder geaccepteerd door de gebruiker.
- c. *Snelheid datamanipulatie (aanwijsvraag, window-leveling en scrollen, zoomen en verplaatsen)*
Deze test is geslaagd als de verschillende handelingen zonder problemen gedaan kunnen worden. Het scrollen door beelden dient vloeiend te verlopen.

Er mag geen groot aantal systemen uitvallen. VQuest kan crashen bij een verkeerd opgestelde toets, maar dat blijkt dan al in de eerste fase van testen. In elke situatie wanneer dit gebeurt, moet dit meteen worden doorgegeven aan het VQuest ontwikkelteam. Het ontwikkelteam kan hiermee de foutenanalyse binnen VQuest verbeteren en daarmee de stabiliteit verhogen.

Ten behoeve van de zekerheid worden de antwoorden op twee plaatsen opgeslagen: op de server en lokaal op de computer. Aan het einde van de toets worden alle lokale antwoorden als back-up weggeschreven naar de server. Bij de toetsen die in de looptijd van het project zijn afgenomen, is het nooit nodig geweest om de back-up van de antwoorden daadwerkelijk te gebruiken. De belasting voor de systemen is echter minimaal, zodat wordt aangeraden dit dubbele systeem wel in stand te houden.

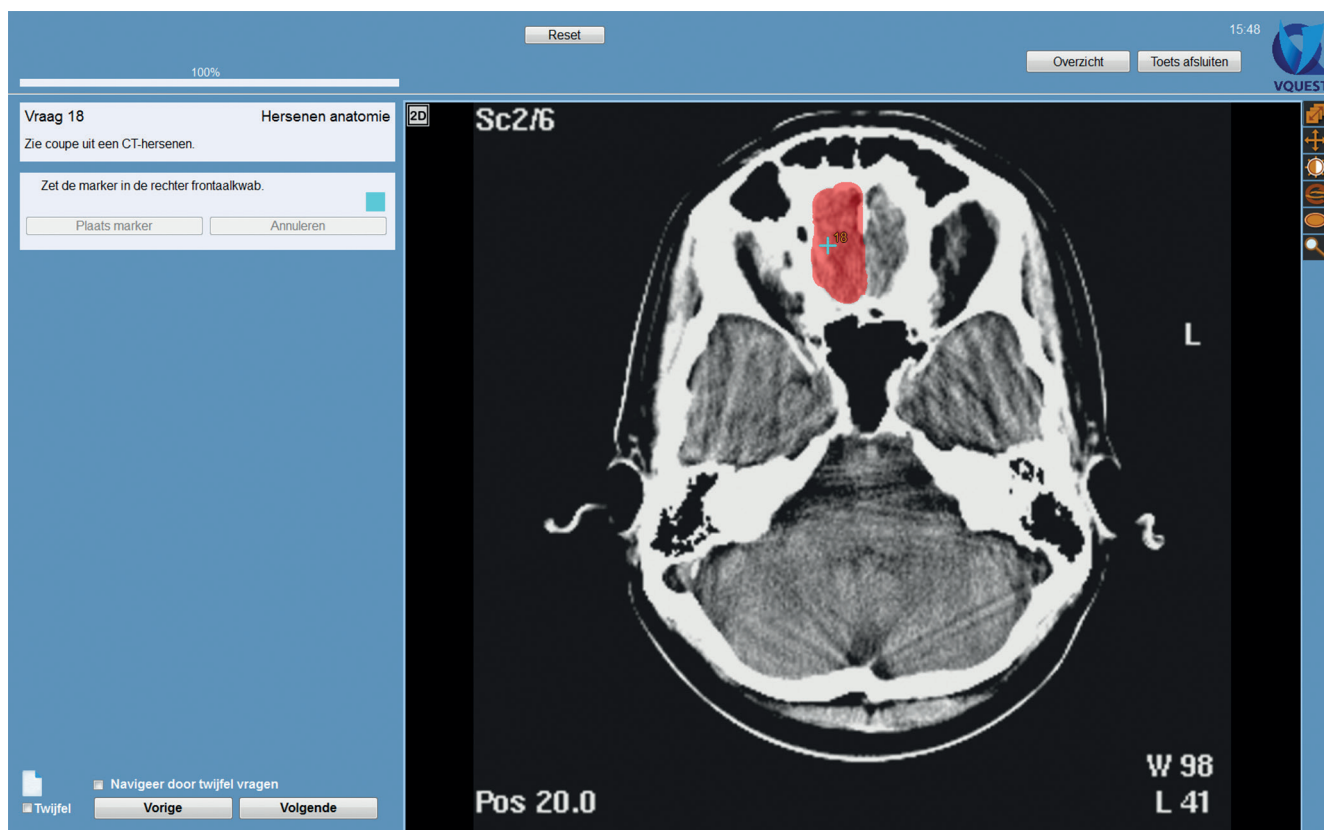
De voortgangstoets radiologie wordt opgesteld door de examencommissie van de Nederlandse Vereniging voor Radiologie (NVvR). Concreet houdt dat in dat het ontwikkelen van de toetsvragen en het samenstellen van de toets halfjaarlijks wordt uitgevoerd door de examencommissie, bestaande uit acht experts in het vakgebied radiologie. Gedurende het opstellen van de vragen wordt een uitgebreide kwaliteitscontrole uitgevoerd, die bestaat uit een drieledige *review*cyclus voorafgaand aan de toetsafname. Na de toetsafname kunnen op basis van de itemanalyse alsnog aanpassingen gedaan worden in het antwoordmodel of toetsvragen komen te vervallen.



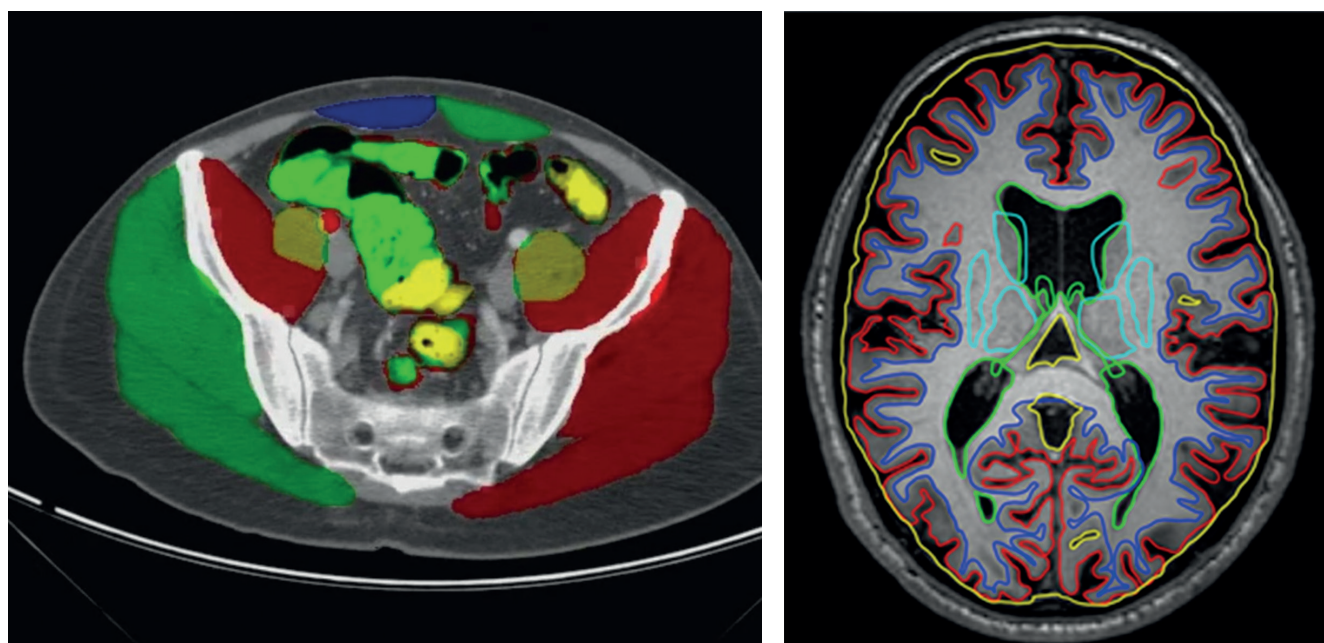
Handreikingen voor implementatie van digitaal toetsen met beelden

flow is dat het dataverkeer tussen examencommissieleden en projectleden zo eenvoudig en efficiënt mogelijk plaats kan vinden. De nieuwe workflow is schematisch weergegeven in Figuur 1c. Na de projectperiode zullen de taken overgenomen worden door de staande organisatie. Het VQuest team zal de digitalisering van de vragen blijven faciliteren.

Het verdient aanbeveling om 1 of 2 mensen verantwoordelijk te maken voor de beeldvragen. Zij verzamelen alle beelddata, zetten de vragen in VQuest en bespreken deze tijdens een gezamenlijke *review*-bijeenkomst. Op afzienbare termijn zal het mogelijk zijn om met VQuest vragen op afstand te delen en te reviewen. Hierbij kan vanaf elke werkplek een verbinding gemaakt worden met een VQuest-server.



Figuur 1d. Voorbeeld van een antwoordmodel (rode structuur).



Figuur 1e. Voorbeelden van twee tekenpakketten om een antwoordmodel te maken.

1.6 Nakijkmodel

Het nakijken van de antwoorden op alle vragen uit een toets kan volledig automatisch en is geïntegreerd in VQuest. Voor de meeste vraagtypen is dit niet anders dan bij andere toetsprogramma's. Het automatisch nakijken van een aanwijsvraag vergt echter wat meer technische voorwaarden dan bij andere vraagtypen. Voor dit vraagtype is het noodzakelijk dat het juiste antwoord wordt ingekleurd door de docent. Dat wordt een overlay genoemd (antwoordmodel), zie Figuur 1d. In het geval van driedimensionale beelden betreft dit dus een volume.

Het maken van een antwoordmodel voor een beeldvraag vergt derhalve een tekenprogramma waarmee driedimensionale beelden kunnen worden ingekleurd. Deze module zal op termijn in VQuest komen.

Voorbeelden van lokaal ontwikkelde intekenprogramma's die dit kunnen staan in Figuur 1e. In het linker beeld van Figuur 1e is een programma gebruikt waarbij op pixelniveau wordt gekleurd. Elke gekleurde pixel geeft een bepaalde structuur weer en wordt als correct beschouwd. In het rechter beeld van Figuur 1e is een contour-gebaseerd programma gebruikt. Daarbij tekent de docent een vloeiende lijn om de juiste structuur. Om dit als antwoordmodel te kunnen gebruiken, moet het uiteindelijke resultaat omgezet worden in een digitaal beeld, waarbij alle pixels een waarde hebben (goed of fout). Deze conversie zit in het tekenpakket.

Hoofdstuk 2. Ontwerpen van een digitale radiologietoets

2.1 Inleiding

Dit hoofdstuk geeft een overzicht van de stappen die nodig zijn om de schriftelijke voortgangstoetsing in de radiologie met 2D-beeldvragen te veranderen naar een digitale voortgangstoets met 2D- en MPR-beeldvragen. Met een herontwerp is het mogelijk om de psychometrische kwaliteit en efficiëntie van de toetsing van radiologische beeldinterpretatievaardigheden te verbeteren. Het herontwerp betekent onder andere dat een structuur voor een digitale itembank moet worden ontworpen. Daarnaast dient het arsenaal aan vraagtypen te worden uitgebreid en dienen mogelijkheden te worden onderzocht voor het samenstellen van een nieuwe toets aan de hand van een itembank en het ontwikkelen van nieuwe scoringsregels en een nakijkmodule. De ontwerpstappen worden vanuit onderwijskundig oogpunt benaderd. De onderwijskundige terminologie is uitgelegd in de begrippenlijst.

Het ontwerpproces, gehanteerd tijdens het project, is geïnspireerd op het evidence-centered design van Mislevy (Mislevy et al. 2004), zie Tabel 2a. Stap 1 behelst het formuleren van het construct beeldinterpretatievaardigheden (studentmodel) en is deels ontwikkeld op basis van een literatuurstudie en empirisch onderzoek in het SURF-project 'Toetsing in beeld' (van der Gijp et al. 2014). In stap 2 zijn de toetstaken en de toetsomgeving vastgesteld (taakmodel). Daartoe zijn een literatuurstudie en toetsanalyses uitgevoerd en heeft een expertsessie met de landelijke examencommissie van de NVvR plaatsgevonden. In stap 3 is het scoringsmodel bepaald (bewijsmodel). Het scoringsmodel is vastgesteld door de examencommissie, na een advies van de projectgroep en een onderwijskundige. In stap 4 vond de samenstelling van de toetsinhoud plaats (de assemblage). De examencommissieleden hebben een training in het samenstellen van de toets gekregen. Tenslotte is de toetsinhoud op digitale wijze vormgegeven in het programma VQuest (presentatie). Dit hoofdstuk beschrijft de ontwerpstappen 1 tot en met 4 zoals doorlopen tijdens het project. Stap 5 betreft de digitalisering van de toets. Randvoorwaarden voor digitalisering worden beschreven in hoofdstuk 1 en de implementatie van de digitale toets wordt beschreven in hoofdstuk 4 van dit handboek.

2.2 Studentmodel: wat behelst het construct beeldinterpretatievaardigheid?

In het studentmodel wordt het te meten construct beeldinterpretatievaardigheid vastgesteld en wordt beschreven hoe dat construct zich ontwikkelt. De vraag die beantwoord moet worden is: Wat zijn beeldinterpretatievaardigheden? Beoordeling van beelden is fundamenteel voor radiologische expertise. In het algemeen kunnen in het beeldinterpretatieproces drie belangrijke stadia worden onderscheiden; 'perceptie', 'analyse' en 'synthese' (Rogers 1996; van der Gijp et al. 2014). De radioloog moet de radiologische afwijkingen eerst verzamelen door ze te zien en te herkennen (perceptie). Vervolgens moeten deze bevindingen uiteengerafeld worden om ze te kunnen begrijpen en beoordelen: wat zijn de kenmerken van de bevindingen/afwijkingen of welke kenmerken ontbreken juist (analyse)? Tot slot moeten deze bevindingen geïntegreerd worden en vertaald worden in een differentiaal diagnose en dient hierop geanticipeerd te worden (synthese). Een differentiaal diagnose is een lijst met diagnosen op volgorde van waarschijnlijkheid op basis van de bevindingen bij het beoordelen van het radiologisch onderzoek.

De radiologische praktijk

In de praktijk beoordeelt de radioloog een radiologisch onderzoek, dat door de behandelend arts van de patiënt is aangevraagd. De aanvragend arts levert hiervoor een vraagstelling in, die op de radiologische beeldvorming van toepassing is, met daarbij de relevante klinische gegevens van de patiënt (bijvoorbeeld symptomen, bevindingen bij lichamelijk onderzoek en/of laboratoriumwaarden). De radioloog beoordeelt het onderzoek (bijvoorbeeld een CT-scan, MRI-scan of een röntgenfoto). Hierbij tracht hij/zij te komen tot een (differentiaal) diagnose. De radioloog maakt een verslag van zijn bevindingen met eventuele aanbevelingen. Een aanbeveling kan bijvoorbeeld zijn dat er vervolgonderzoek plaats moet vinden.

Tabel 2a. Overzicht toetsherontwikkeling geïnspireerd op Evidence-centered design

Stap in ECD*	Kernvraag	Methode	Opbrengst
1. Studentmodel	Wat houden beeldinterpretatievaardigheden in? Wat willen we toetsen?	Literatuurstudie en empirisch onderzoek in het SURF-project 'Toetsing in beeld'	- Beschrijving van het construct en de ontwikkeling van beeldinterpretatievaardigheden
2. Taakmodel	Welke (toets)vragen (zowel inhoudelijk als qua vorm) zijn nodig om informatie te krijgen over de beeldinterpretatievaardigheden van de aios? Welke essentiële taakkenmerken moet het toetsprogramma (VQuest) daarvoor hebben?	Literatuurstudie Item en toetsanalyse op bestaande data Expertsessie met examencommissie	- Een toetsmatrijs met te dekken onderwerpen en vaardigheden in de toetsing van beeldinterpretatievaardigheden. - Vraagtypen die geschikt zijn om beeldinterpretatievaardigheden te meten.
3. Bewijsmodel	Wat telt als bewijs voor beeldinterpretatievaardigheden en hoe interpreteren we dit bewijs?	Item- en toetsanalyse op bestaande data Standaard setting methode	- Scoringsregels, nakijkmodel en cesuurbepaling van de voortgangstoets radiologie met 2D- en MPR-vragen. - Een samengestelde voortgangstoets met 2D- en MPR-beeldvragen.
4. Assemblage	Hoeveel taken en welke soort taken zijn nodig in de complete toets? (welke vragen en in welke volgorde van moeilijkheid etc.)	Training en toetsconstructie met leden van de examencommissie radiologie	- Een training over toetsconstructie. - Een basis van een digitale itembank met 2D- en MPR-beeldvragen, waarin in ieder geval voldoende beeldvragen voor de eerste landelijke digitale voortgangstoetsen opgenomen zijn. - De toekomstige digitale itembank bevat een database waarin kwaliteitsindicatoren per vraag kunnen worden opgenomen na toetsafname.
5. Presentatie	Hoe wordt de toets gepresenteerd, hoe ziet de toets eruit? Wat zou de afnamesituatie moeten zijn? Wat zijn technische, materiële voorwaarden om taakuitvoering op te roepen? Hoe moeten de scores weergegeven worden?	Gebruik maken van ervaringen tijdens pilots Gebruik maken van gegevens uit de literatuur	- Een gebruiksvriendelijk en intuïtief interface voor toetsafname en toetsconstructie (H1, Figuur 1b). - Specificaties over een overzichtelijke en effectieve manier van score representatie (H3).

*ECD = Evidence-centered design (Mislevy et al. 2004) .

De laatste jaren is de dagelijkse praktijk van beeldinterpretatie veranderd als gevolg van de digitalisering en de introductie en ontwikkeling van nieuwe radiologische technieken. Hierdoor hebben ook MPR-beelden hun intrede in de radiologische praktijk genomen. Omdat de beoordeling van MPR-beelden om andere perceptuele en cognitieve activiteiten vraagt, vergeleken met beoordeling van 2D-beelden (Hegarty 2010), is er een verschuiving ontstaan in de benodigde radiologische expertise. Om de expertise van radiologen in de opleiding te garanderen is het van groot belang om hun training en toetsing bij deze ontwikkelingen aan te laten sluiten en daarop niet achter te laten lopen. Om dit te kunnen uitvoeren is het belangrijk vast te stellen uit welke elementen beeldinterpretatievaardigheden in de huidige radiologische praktijk bestaan.

Een literatuurstudie naar elementen van beeldinterpretatievaardigheden is verricht om hier inzicht in te krijgen (van der Gijp et al. 2014). De studie werd uitgevoerd middels een uitgebreide search in de databases PubMed, PsycINFO en Embase, aangevuld met de referenties van de artikelen en relevante hoofdstukken in boeken over medische beeldvorming (range jaartal onbepaald). Uit de literatuurstu-

die blijkt dat er twee stromingen bestaan: een perceptuele en een cognitieve stroming (Beutel 2000), welke voornamelijk gebaseerd zijn op respectievelijk oogbeweging (eye-tracking) studies (Cooper et al. 2009; Krupinski 1996; Kundel and La Follette 1972; Kundel et al. 1978; Kundel et al. 2007; Nodine and Kundel 1987a; Nodine et al. 2002) en studies die het hardop denken tijdens het interpreteren van beelden analyseren (verbale protocollen) (Azevedo et al. 1997; Lesgold et al. 1988; Morita et al. 2008; Raufaste et al. 1998; Rogers 1996). Eye-tracking studies laten zien dat verschillen in beeldinterpretatievaardigheden berusten op perceptuele activiteiten (zoals zoekstrategieën en fixaties van de ogen). Hardopdenkstudies rapporteren verschillen in cognitieve activiteiten, zoals redeneringen, evaluatie van radiologische bevindingen en het gebruik van verschillende soorten kennis (bijvoorbeeld anatomie en techniek). Helaas richtten de studies zich tot nu toe op de beoordeling van conventionele röntgenfoto's (2D-beelden) en zijn er nog maar weinig studies verricht naar de vaardigheden die nodig zijn voor het beoordelen van MPR-beelden.

De basis van het studentmodel

In het geval van beeldinterpretatie kunnen drie fasen worden onderscheiden: perceptie, analyse en synthese (Kundel et al. 1978; Nodine and Kundel 1987b; Rogers 1996; van der Gijp et al. 2014). Deze fasen vinden alternerend plaats tijdens het beeldinterpretatieproces. Tijdens deze drie fasen worden zowel perceptuele als cognitieve vaardigheden aangesproken. Deze vaardigheden kunnen worden ingedeeld op basis van de taxonomie van Bloom (Bloom 1956): kennis, begrip, toepassen, analyseren, evalueren en synthetiseren.

De drie componenten van radiologische beeldinterpretatie

Perceptie

Om een afwijking te kunnen beoordelen, moet de afwijking eerst herkend worden als afwijkend. Dit herkennen van afwijkingen speelt zich af in de perceptiefase. Hiervoor is het essentieel te weten hoe een normaal beeld eruit ziet. De radioloog in opleiding heeft dus kennis en begrip nodig over de normale anatomie. Deze kennis moet vervolgens toegepast worden op de betreffende casus. Welke structuren zijn in dit geval normaal en welke afwijkend?

Analyse

Indien een afwijking geïdentificeerd is, dient deze geanalyseerd te worden. Een afwijking kan bijvoorbeeld scherp begrensd zijn of juist glad begrensd, groot of klein, en heel donker of juist licht weergegeven. Dit zegt iets over de oorzaak van de afwijking. Ook de locatie van de afwijking kan richtinggevend zijn. Het analyseren van afwijkingen vergt kennis en begrip over de verschillende beeldkarakteristieken van afwijkingen, die door ziekteprocessen veroorzaakt worden en over de afbeeldingstechniek die gebruikt is.

Synthese

Zoals eerder beschreven, is het uiteindelijk van belang tot een (differentiaal) diagnose te komen. Op basis van de beeldkarakteristieken van de afwijking(en) in combinatie met de patiëntgegevens komt de radioloog tot verschillende opties van diagnoses. Deze zet hij/zij in volgorde van waarschijnlijkheid. Bij deze rangorde bepaling worden niet alleen informatie opgedaan tijdens het analyseproces gebruikt, maar ook allerlei informatie die de radioloog eerder heeft verkregen over de waarschijnlijkheid van de aan- of afwezigheid van ziekten op andere gronden. Zo kunnen bijvoorbeeld ook epidemiologische factoren een rol spelen, omdat bepaalde ziektebeelden bij bepaalde (leeftijds)groepen vaker of juist minder vaak voorkomen. De radioloog synthetiseert de differentiaal diagnose dus op basis van de opgedane informatie in combinatie met kennis en begrip (bijvoorbeeld over epidemiologie en over ziekteprocessen).

Conclusie

Voor het kunnen interpreteren van radiologische beelden zijn zowel perceptuele als cognitieve vaardigheden essentieel. De afwijkingen moeten herkend worden (perceptie), de afwijkingen moeten gekarak-

teriseerd worden (analyse) en vervolgens dient een differentiaal diagnose opgesteld te worden, indien mogelijk (synthese). In dit beeldinterpretatieproces zijn de verschillende niveaus van de taxonomie van Bloom te herkennen. De perceptuele en cognitieve vaardigheden die nodig zijn voor beeldinterpretatie zijn opgenomen in een schematisch weergegeven studentmodel. Dit model is gepubliceerd in hoofdstuk 2.1 van het handboek 'Digitale toetsing van radiologische beeldinterpretatie' (Ravesloot 2014).

2.3 Taakmodel: welke toetsvragen zijn inhoudelijk en vormtechnisch relevant voor het toetsen van beeldinterpretatievaardigheden?

2.3.1. Toetsomgeving

Toelichting

Het taakmodel betreft de keuze van essentiële toetstaken en de exacte toetsomgeving die nodig zijn om te werken met MPR-beelden om zodoende informatie te krijgen over de beeldinterpretatievaardigheden van een aios radiologie. Het bevat tevens specificeringen van het type benodigde toetsomgeving, bijvoorbeeld kenmerken van het stimulusmateriaal (type vraag, plaatjes, interface), gegeven instructies etc. Het taakmodel voor de toetsing van beeldinterpretatievaardigheden is reeds deels ontwikkeld in het programma VQuest (Vincken & Ravesloot, 2010). Zo is de kern van de userinterface altijd een afbeelding van een 2D- of MPR-beeld, waarvan de MPR-beelden in meerdere richtingen bekeken kunnen worden, zoals opleiders en aios gewend zijn als ze achter een computer zitten in een klinische setting. Op de site www.vquest.nl kunt u meer informatie vinden over de toetsomgeving VQuest (bijvoorbeeld door de instructiefilm te bekijken).

2.3.2 Toetsmatrijs

Toelichting

Essentieel voor de ontwikkeling van het taakmodel is een toetsmatrijs. In de toetsmatrijs worden de leerdoelen en eindtermen van de opleiding radiologie in de toetsing vastgelegd. Met het oog op de inhouds- en constructvaliditeit is een evenwichtige en representatieve dekking van eindtermen essentieel, evenals goedkeuring daarvan door experts van de landelijke examencommissie van de Nederlandse Vereniging voor Radiologie.

Methode

De toetsmatrijs is opgesteld op basis van de leerdoelen uit het HORA-opleidingsplan. Dit is het opleidingsplan van de specialisatie radiologie, dat in 2010 is herzien in verband met veranderingen in de opleidingsstructuur. De leerdoelen zijn per deelgebied beschreven. De genoemde ziektebeelden kunnen bevraagd worden met behulp van tekstvragen, voornamelijk gericht op kennis over het ziektebeeld, of met behulp van beeldvragen, vooral gericht op het toepassen van deze kennis voor het detecteren van afwijkingen en het diagnosticeren van ziektebeelden. Voor alle leerdoelen in het HORA-opleidingsplan is door de projectgroep aangegeven welke modaliteit in de praktijk doorgaans gebruikt wordt voor het in beeld brengen van de afwijkingen/de ziektebeelden. In Tabel 2b staat een voorbeeld hiervan. Hierbij zijn echo en doorlichtonderzoeken achterwege gelaten, omdat dit dynamische onderzoeken zijn. Hierbij gaat het niet alleen om de vaardigheid van het beoordelen van de beelden, maar ook het vervaardigen van de beelden zelf (beeldacquisitie). De beeldacquisitie en beoordeling van deze beelden zijn daarom vooralsnog het meest geschikt voor toetsing in de praktijk.

Tabel 2b. Voorbeeld van leerdoelen thoraxradiologie met hieraan gekoppelde belangrijke modaliteiten

Leerdoelen: anatomie/ afwijkingen/ziektebeelden <i>Bron: HORA-opleidingsplan 2010</i>	Modaliteit		
	<i>Röntgen</i>	<i>CT</i>	<i>MR</i>
Anatomie	x	x	
Interstitiële afwijkingen	x	x	
Alveolaire afwijkingen	x	x	
Cysten/caviteiten	x	x	
Intensive care	x	x	
Abnormale luchtcollecties	x	x	
Chronisch diffuse + obstructieve longafwijkingen	x	x	
Infecties + complicaties - o AIDS - o Immunologisch	x	x	
Pulmonale verdichtingen	x	x	
Trauma	x	x	
Neoplasmata	x	x	
Pleurale en diafragma afwijkingen	x	x	
Congenitale longafwijkingen	x	x	
Vasculaire afwijkingen	x	x	x

Legenda: CT = Computed Tomography; MR = Magnetic Resonance

Opbrengst

De toetsmatrijs voor de digitale VGT bouwt voort op de toetsmatrijs van de huidige VGT, welke door de examencommissie is samengesteld op basis van de leerdoelen uit het HORA-opleidingsplan. De toets bestaat uit 10 deelgebieden. Het aandeel per deelgebied in de toets hangt samen met het aandeel van het deelgebied in de opleiding en de dagelijkse praktijk. Er worden bijvoorbeeld relatief veel thoraxfoto's en CT scans van de thorax gemaakt in de praktijk en daarom heeft dit een groot aandeel in de toets. Ook het aantal beeldvragen per deelgebied is door de examencommissie vastgesteld. Voor de digitale VGT wordt het aantal beeldvragen in eerste instantie gelijk gehouden, maar in de toekomst zal de verhouding beeldvragen/tekstvragen mogelijk veranderen. Dit hangt af van meerdere factoren. Ten eerste moet de kwaliteit van de nieuwe digitale beeldvragen (zowel 2D als MPR) eerst vastgesteld worden op basis van de item- en toetsanalyses van de eerste digitale voortgangstoetsen. Daarnaast is het natuurlijk van belang dat het technisch haalbaar is om het aantal (MPR) beeldvragen op te hogen in verband met de dataload van de toets. In de eerste digitale VGT is als pilot één MPR-vraag per deelgebied opgenomen. Aangezien dit succesvol is verlopen is het aantal MPR-vragen in de daaropvolgende twee digitale VGT's uitgebreid (21 en 16 MPR-beeldvragen). In onderstaande toetsmatrijs (Tabel 2c) is een voorstel gedaan voor de verdeling van de modaliteiten (X, CT of MR) van de beeldvragen per deelgebied, waarbij het aantal beeldvragen voornamelijk dus gelijk is gehouden. De aantallen per modaliteit zijn gebaseerd op de leerdoelen. Per leerdoel is aangegeven welke modaliteiten in de praktijk gebruikt worden om het ziektebeeld in beeld te brengen. Voor bijvoorbeeld het hoofd-hals gebied en de hersenen wordt hoofdzakelijk gebruik gemaakt van MRI en CT, terwijl CT juist weer minder geschikt is voor mammariadiologie en MRI minder geschikt voor thoraxradiologie. Deze verdeling van modaliteiten in de leerdoelenlijst is in de toetsmatrijs verwerkt. Sommige deelgebieden hebben slechts een klein aantal beeldvragen, terwijl wel meerdere modaliteiten gebruikt worden voor de beeldvorming in dit deelgebied. In deze gevallen kan door de vragenmakers gekozen worden voor één of twee van de modaliteiten.

Tabel 2c. Voorbeeld toetsmatrijs voor digitale toetsing in de radiologie

Deelgebied	Aantal vragen	Aantal tekstvragen	Aantal X	Aantal CT	Aantal MR
Thorax	29	22	4	3	0
Hoofd-hals	14	12	0	1	1
Musculoskeletaal	30	24	2	2	2
Cardiovasculair	11	9	0-1	0-1	0-1
Abdomen (gastro-intestinaal)	29	23	0	4	2
Urogenitaal	18	14	1	2-3	0-1
Kinderradiologie	20	19	0-1	0-1	0-1
Neuroradiologie	30	24	0	3	3
Mammaradiologie	10	8	1	0	1
Interventie	9	8	0-1	0-1	0-1

Het vakgebied radiologie is voortdurend in ontwikkeling. Dit geldt ook voor de opleiding tot radioloog. In de afgelopen jaren is bijvoorbeeld veel vooruitgang geboekt op het gebied van beeldvorming van het hart. Dit heeft geleid tot een nieuw deelgebied; cardiovasculair. Bovenstaande toetsmatrijs is dan ook dynamisch. Het kan zijn dat bepaalde deelgebieden of modaliteiten in de toekomst belangrijker of juist minder belangrijk zullen worden, en dat de toetsmatrijs hierop aangepast zal worden.

2.3.3 Soort toetsvragen

Toelichting

Tevens bevat het taakmodel het soort toetsvragen dat gesteld gaat worden. Voorafgaand aan dit project werd in radiologische voortgangstoetsing enkel gewerkt met 'juist/onjuist/weet-niet vragen'. Men kan zich afvragen of dit typen vraag voldoende relevant is voor het te meten construct. Daarom zijn in het project ook andersoortige vragen dan voorheen (in de schriftelijke voortgangstoetsing) gebruikelijk was onderzocht binnen het programma VQuest. Het gaat daarbij om de volgende vraagtypen:

1. Aanwijsvraag: een type hotspotvraag waarbij de kandidaat een marker binnen een juiste structuur moet zetten;
2. Pijlvraag; de kandidaat moet een pijl in de juiste richting en eventueel ook nog in een bepaalde structuur zetten;
3. Meetlatvraag waarbij de kandidaat een structuur met een meetlatje moet opmeten.
4. Multiple choice vraag (MC-vraag)
5. Meer-uit-meer vraag (MM-vraag)
6. Dropdownvraag: een type longmenuvraag die raakt aan een open vraag, waarbij de kandidaat een antwoord moet kiezen uit een lijst met zeer veel opties (liefst honderden) die verschijnt door het intypen van de eerste letters van het antwoord (het aantal letters is in te stellen naar keuze). Ook antwoorden waarin deze lettercombinatie niet aan het begin voorkomt zullen verschijnen, bij het intypen van de lettercombinatie blo verschijnt dus ook het antwoord: her-senbloeding.

Onderdelen van beeldinterpretatie zijn perceptie, analyse en synthese. Voor het toetsen van perceptie zou de aanwijsvraag een geschikt vraagtype kunnen zijn. Voor het analyseren van afwijkingen of het opstellen van een differentiaal diagnose (onderdeel van synthese) kan gedacht worden aan een meer-uit-meer vraag, een multiple choice vraag of een dropdown menu vraag. Uiteindelijk wordt beoogd om een variatie aan vraagsoorten te creëren die de basis vormen voor de te ontwikkelen itembank.

Methode

De mogelijke vraagtypen zijn ingebracht in een expertbijeenkomst met leden van de examencommissie radiologie. Op basis hiervan is een selectie van geschikte vraagtypen gemaakt. De aanwijsvraag, MC-vraag, MM-vraag en dropdownvraag werden door de examencommissie geschikt bevonden. In een gerichte training over itemconstructie in VQuest werd behandeld voor welke toetsdoelen deze vraagtypen het best inzetbaar zijn en hoe deze vraagtypen geconstrueerd kunnen worden binnen de ontwikkelde toetsmatrijs.

Opbrengst

De opbrengst zijn de uiteindelijke toetsvragen in de toets. Deze zijn niet in dit rapport opgenomen.

2.4 Bewijsmodel: wat telt als bewijs voor beeldinterpretatievaardigheid?

Toelichting

Wat telt als bewijs voor het beheersingsniveau van beeldinterpretatievaardigheden van kandidaten en hoe interpreteren we dit bewijs? Hierbij is gebruik gemaakt van de resultaten die zijn verkregen door analyses van de toetsscores van de beeldvragen in het SURF-project 'Toetsing in beeld'. In dat onderzoek is ook de gemiddelde responsetijd per item (2D- of 3D-beeldvraag) vastgesteld. Deze gegevens zijn benut om te bepalen of en in welke mate de bestaande scoringsregels en antwoordmodellen in de voortgangstoets radiologie moeten worden aangepast. In het geval van de voortgangstoets radiologie werd tot voor kort gewerkt met formula scoring. Dit betekent dat juist/onjuist-vragen een weet-niet optie bevatten en scores bepaald worden door het aantal foute antwoorden af te trekken van het aantal goede antwoorden (goed min fout score). Voor het antwoord weet-niet worden geen punten gerekend. Echter bij uitbreiding van vraagtypen (anders dan juist/onjuist/weet-niet) kon daarmee niet meer worden volstaan en zijn nieuwe scoringsregels en antwoordmodellen ontwikkeld. Op basis van de item- en toetsanalyses zal de optimale toetslengte van de MPR-beeld toetsing worden bepaald. Het bepalen daarvan vergt een afweging van de kwaliteit van de items en toets, de mogelijke afnametijd, kenmerken van de kandidaten en het te dekken inhoudsdomen met de te meten vaardigheden. De cesuurbepaling (het vaststellen van een cesuur die de grens aangeeft tussen voldoende en onvoldoende) is norm-gerefereerd.

Methode

Een item- en toetsanalyse op bestaande toetsgegevens van de voortgangstoets radiologie waarin is gewerkt met 2D- en MPR-beeldvragen zijn uitgevoerd. Dit betreft de scores van ruim 1100 afgelegde voortgangstoetsen radiologie in Nederland. Ten eerste is een itemanalyse uitgevoerd, waaronder het berekenen van rit-waarden en p-waarden. Op basis van de analyses is door de examencommissie besloten welke items uit de toets moeten worden verwijderd. Ten tweede is een toetsanalyse uitgevoerd op basis waarvan informatie wordt verkregen over de toetsscores in de gehele groep aios. Hiervoor worden een betrouwbaarheidsanalyse (t.b.v. het in kaart brengen van de betrouwbaarheid) uitgevoerd. Daarna worden de individuele toetsresultaten berekend, het betreft hier een relatieve normering (McHarg et al. 2005), dat wil zeggen, een score ten opzichte van het hele cohort en een score ten opzichte van de kandidaten met gelijke opleidingservaring.

Opbrengst

De resultaten van de toets- en itemanalyse worden besproken in Hoofdstuk 5a.

2.5 Assemblage: toetssamenstelling

Toelichting

De assemblage betreft de eigenlijke toetssamenstelling waarin de bevindingen van alle voorafgaande stappen worden samengevoegd. Hierbij wordt de toets definitief samengesteld en worden beslissingen genomen over hoeveel items en welk soort items in welke volgorde nodig zijn in de complete toets. Acht deelcommissieleden en examencommissieleden zijn verantwoordelijk voor de uitvoering van deze stap en zijn hiervoor getraind.

Methode

Het welslagen van het toetsen van beeldinterpretatievaardigheden staat of valt grotendeels met de kwaliteiten van de toetsconstructeurs. Bij deze stap is dan ook een efficiënte training voor toetsconstructie ontwikkeld en geïmplementeerd om toetsconstructeurs een toets met het programma VQuest te leren samenstellen. Daarin is de gehele procedure van het werken als docent met VQuest doorlopen en geoefend. De volgende onderwerpen zijn aan bod gekomen:

- het maken van toetsvragen in het programma VQuest gebaseerd op leerdoelen en eindtermen van de opleiding radiologie.
- de workflow voor het maken en reviewen van de digitale toetsvragen.
- het gebruik van beeldbewerkingsfuncties in VQuest.

De training is opgezet volgens het principe van *blended learning* en bestond uit een huiswerkopdracht (deelnemers dienden digitale beeldvragen in), een groepsbijeenkomst van 2 uur (de training), bestaande uit een plenair gedeelte en een hands-on gedeelte waarin de gemaakte digitale vragen in VQuest in groepjes van twee werden bestudeerd onder individuele begeleiding. Om de effectiviteit van de training te vergroten is de training geaccrediteerd en afgesloten met het ontwikkelen van een deel van een voortgangstoets die, bij voldoende kwaliteit, daadwerkelijk als onderdeel van de voortgangstoets binnen de radiologie werd ingezet.

Opbrengst

De eerste digitale voortgangstoets bestond uit 21 2D-vragen en 11 MPR-vragen, alle juist/onjuist/weet-niet vragen. De tweede digitale voortgangstoets bestond uit 16 2D-vragen en 21 MPR-vragen met variërende vraagtypen: aanwijs, MC, MM en juist/onjuist.

Hoofdstuk 3. Feedback bij toetsing van radiologische beeldinterpretatie

3.1 Inleiding

Radiologische beeldinterpretatie is een complexe vaardigheid, waarin kennis en vaardigheden met elkaar interacteren voor het extraheren van informatie uit het beeld en het geven van betekenis aan deze (visuele) informatie (van der Gijp et al. 2014). Voor het opsporen van de relevante visuele informatie is efficiënt zoeken van belang. Bij de beoordeling van MPR-beelden heeft de radioloog een grote hoeveelheid beeldinformatie ter beschikking, waarbij in meerdere richtingen door het beeld heen gescrold kan worden en waarbij het contrastvenster kan worden aangepast. Keuzes voor de juiste beeldbewerkingsmethoden zijn essentieel om efficiënt de informatie te vinden. Radiologische beeldinterpretatie is logischerwijs complex en vereist een lang leertraject. Effectieve feedback is cruciaal om de ontwikkeling van beeldinterpretatievaardigheden te stimuleren. Dit is echter geen sinecure. Want, hoewel de performance van kandidaten op toetsing van beeldinterpretatievaardigheden eenvoudig in kaart is te brengen (bijvoorbeeld het aantal goede scores op een toets), is het onderliggende proces van beeldinterpretatie een black box. Inzicht hierin is nodig om feedbackmogelijkheden op zowel de performance als het proces van beeldinterpretatie te verbeteren.

Onder feedback verstaan we informatie die wordt gecommuniceerd naar kandidaten met als doel hun denken of doen te beïnvloeden opdat het leren en de prestatie verbeteren (Shute 2008). Goede feedback is efficiënt en constructief. De feedback kan bij voorkeur ‘snel’ worden gegeven. Feedback kan kandidaten ondersteunen door aan te geven of een gegeven antwoord goed of niet goed is (confirmerende of corrigerende feedback). Feedback heeft echter niet alleen als doel de score terug te koppelen, maar ook om het leerproces van kandidaten te verbeteren. Goede feedback geeft dan ook niet alleen aan dat een antwoord goed of fout is, maar bevat ook suggesties voor verbetering (Sadler 1998). Idealiter zorgt de feedback dat de student: (1) een idee heeft van de te behalen norm (een idee hebben van wat goed is en wat fout); (2) zijn huidige niveau kan vergelijken met deze norm – dit betekent dat hij in staat is zijn eigen handelen te evalueren en te monitoren; en (3) kan overgaan tot acties en tactieken om de eventuele kloof tussen de gewenste norm en het behaalde niveau van handelen te overbruggen (Sadler 1989; Sadler 1998; Sadler 2010).

Feedback kan mondeling, schriftelijk of digitaal gegeven worden. Digitale feedback kan vaak automatisch gegenereerd worden en is daardoor efficiënt. Docenten en kandidaten kunnen op deze manier snel (just-in-time) inzicht krijgen in wat de kandidaten goed en minder goed doen en waardoor dat wordt veroorzaakt. Tevens biedt digitale feedback mogelijkheden voor het geven van (complexe) visuele feedback (bijvoorbeeld grafieken) of dynamische visuele feedback (bijvoorbeeld met behulp van filmpjes).

In dit hoofdstuk worden feedbackmogelijkheden van digitale feedback bij het toetsen van beeldinterpretatie in de radiologie aan de hand van het programma VQuest beschreven. Ten eerste wordt het geven van feedback op toetsscores behandeld. Daarbij wordt gebruik gemaakt van een voorbeeld van de mogelijkheden tot directe feedback in het toetsprogramma VQuest. Ten slotte volgt een beschrijving met voorbeelden over feedback op het proces van beeldinterpretatie.

3.2 Feedback geven op behaalde scores

Voorwaardelijk voor efficiënte feedback is dat het ‘snel’ kan worden gegeven. In VQuest is hiervoor een module ontwikkeld. Alle vraagtypen in VQuest kunnen automatisch worden nagekeken. De juiste antwoorden zijn door de docent ingevoerd bij het maken van de vragen. Voor de aanwijsvragen bestaat het antwoordmodel uit één of meer ingekleurde structuren. Deze worden in VQuest weergegeven als een overlay over de beelddata. Tijdens het inkijken van een toets is deze overlay wel zichtbaar, alsmede de positie van de markering. De uitvoer van de nakijkmodule heeft meerdere niveaus van informatie.

In het simpelste geval, bestaat de uitvoer uitsluitend uit het aantal punten per vraag per kandidaat. De uitgebreide versie bevat enkele kolommen per vraag per kandidaat, met onder meer het aantal punten, het juiste antwoord, en het gegeven antwoord. Met deze informatie kan de docent eenvoudig inzien welke (incorrecte) antwoorden vaak zijn gegeven en kan daar actie op ondernemen (bv antwoordmodel aanpassen of vraag laten vervallen).

De nakijkmodule heeft tevens een optie om een uitvoerfile per kandidaat te genereren met daarin de toetsvragen, de juiste antwoorden en de gegeven antwoorden door de kandidaat. Deze file kan vervolgens naar elke individuele kandidaat gestuurd worden met de voorlopige totaalscore. De kandidaat ontvangt op deze manier kort na de toetsafname de behaalde score, gespecificeerd naar hoe deze score is opgebouwd. Een vereenvoudigd voorbeeld is opgenomen in Figuur 3a. In dit project is deze module op deze manier gebruikt om het voorlopige toetsresultaat aan de aios te mailen. Na het sturen van het voorlopige toetsresultaat was er tijd om een uitgebreide itemanalyse te doen en was het voor de aios mogelijk om commentaar te leveren op de vragen (binnen één week na het voorlopige toetsresultaat). Naar aanleiding hiervan kan de examencommissie alsnog besluiten tot het aanpassen van de antwoordsleutel of elimineren van vragen.

Voor de feedback op de definitieve toetsscores van de kandidaten wordt bij de voortgangstoets radiologie gebruik gemaakt van een uitgebreide rapportage. In deze rapportage wordt de toetsscore vergeleken met de toetsscores van aios uit hetzelfde jaar en met de toetsscores van aios uit het hele cohort. Zo krijgt de kandidaat een indruk van zijn/haar relatieve toetsscore ten opzichte van de gewenste norm (1 en 2). De feedback is zowel momentaan, betrekking hebbend op de scores op de vragen van één toets, als longitudinaal (scoreverloop in de tijd representerend) van aard. Dit is grafisch weergegeven, zodat aios in één oogopslag kunnen zien hoe zij presteren ten opzichte van de jaargroep en hoe de voortgang in de afgelopen jaren is geweest. De scores zijn tevens uitgesplitst naar deelgebied, zodat de aios kan differentiëren waar zijn/haar hiaten of juist kwaliteiten liggen.

In de medische vervolgopleiding is het gebruikelijk dat elke opleidingskliniek een opleider heeft. Dit is één van de radiologen, die hoofdverantwoordelijk is voor de opleiding in zijn kliniek en veel opleidingstaken vervult. De voortgangsgesprekken worden bijvoorbeeld uitgevoerd door de opleider (vaak samen met de vice-opleider). Voor de opleider is het dan ook van belang om informatie te ontvangen over de scores van de aios uit zijn/haar kliniek. Daarom krijgen ook de opleiders een feedback rapportage. Dit bestaat uit de scores per arts-assistent ten opzichte van zijn/haar jaargroep en ten opzicht van het hele cohort. Ook wordt de gemiddelde score van de opleidingskliniek vergeleken met de scores van de andere klinieken. Op deze manier krijgt de opleider ook feedback op zijn opleidingspraktijk en kan indien nodig bijsturen. Bijvoorbeeld als zijn aios relatief laag scoren, kan hij in de individuele prestaties van de aios nagaan of dit komt doordat één bepaald onderwerp niet voldoende wordt beheerst of doordat juist op alle onderwerpen laag wordt gescoord. In het eerste geval kan de opleider in zijn opleidingskliniek meer tijd inruimen voor dat specifieke onderwerp, en in het tweede geval kan hij bijvoorbeeld meer onderwijsmomenten inroosteren voor de aios.

3.3 Feedback geven op het beeldinterpretatieproces

Naast het terugkoppelen van de behaalde toetsscore, kan met het programma VQuest ook feedback worden gegeven op het handelingsproces tijdens het interpreteren van beelden. In het toetsprogramma VQuest zijn meerdere mogelijkheden ontwikkeld voor het geven van feedback op het beeldinterpretatieproces. De wijze waarop deze feedback wordt gegenereerd sluit aan bij het idee van *learning analytics*; het digitale gedrag (klikken met de muis, scrollen etc.) van de kandidaat wordt vastgelegd en gebruikt om feedback te genereren. Deze informatie kan zowel voor visuele feedback als voor tekstuele en getalsmatige feedback gebruikt worden, om het handelingsproces terug te koppelen en om suggesties te geven voor verbetering.

Toets: VGT radiologie, 19 april 2013

AIOS: Anoniem

NVvR nr: 0000

e-mail: anoniem@gmail.com

Hieronder staat een overzicht van alle tekst- en beeldvragen, met zowel het correcte antwoord op de vraag als het door u gegeven antwoord.

Totaal aantal vragen: 200

1. (CV Beeldvraag)

Casus: Een transversale CT-scan door het bovenste mediastinum, blanco en in arteriële fase na IV contrastmiddel.

Vraag: Het beeld past beter bij een dissectie dan bij een aneurysma met een wandstandige thrombus.

Correcte antwoord: Juist

Gegeven antwoord: Juist

2. (CV Beeldvraag)

Casus: De aortaklep heeft 3 cusps.

Vraag: Het sterretje staat caudaal van de non-coronary cusp.

Correcte antwoord: Juist

Gegeven antwoord: Juist

198. (Uro Beeldvraag)

Casus: Retrograad urethrogram bij een man.

Vraag: Er is een stenose in het pars membranosa.

Correcte antwoord: Onjuist

Gegeven antwoord: Juist

199. (Uro Beeldvraag)

Casus: Axiale CT-afbeeldingen vervaardigd bij een 20-jarige patiënte met flankpijn rechts.

Vraag: Bovenaan in de differentiaal diagnose staat een uretersteen rechts.

Correcte antwoord: Onjuist

Gegeven antwoord: Onjuist

200. (Uro Beeldvraag)

Casus: Axiale coupes uit een CT-IVP bij een vrouw van middelbare leeftijd (gemengd nefroge/excretiefase).

Vraag: Rechts is er een stenose op de uretero-pelviene overgang.

Correcte antwoord: Onjuist

Gegeven antwoord: Onjuist

Overzicht:

Totaal aantal vragen beantwoord: 200

Aantal vragen niet bekeken: 0

Aantal vragen GOED beantwoord: 88

Aantal vragen FOUT beantwoord: 23

Aantal vragen met WEET NIET beantwoord: 89

Figuur 3a. Voorbeeld van de uitslagmail met voorlopige totaalscore.

Visuele feedbackmodule

De logfiles kunnen gebruikt worden om het digitale gedrag van de kandidaat opnieuw af te laten spelen. Dit resulteert in een filmpje van de zoekstrategie van de kandidaat. Dit wordt de *playback* module genoemd. Zo kan de kandidaat zien waar hij/zij heeft gezocht tijdens het interpreteren van het beeld en krijgt hij/zij inzicht in het eigen beeldinterpretatie proces. Dit geeft echter nog geen handreikingen voor het verbeteren van de zoekstrategie. Daarvoor is een expertmodule gemaakt. In deze module kunnen docenten een zoekstrategie van een expert (al dan niet de docent zelf) opnemen. Hierbij kan additioneel auditieve informatie over de achterliggende gedachten van de expert om deze zoekstrategie te hanteren, worden opgenomen. Dit laatste is nu nog niet mogelijk in VQuest, maar kan eenvoudig met een separaat programma worden opgenomen. Deze expertfilms geven de kandidaat handreikingen voor het aanleren van nieuwe, mogelijk efficiëntere zoekstrategieën. Een eerste studie verricht in het kader van het aanverwante SURF-project IMAGE ME (Vincken et al. 2013) naar de effectiviteit van deze

nieuwe feedback methoden laat zien dat de studenten zelf denken baat te hebben bij deze aanvullende feedback over het handelingsproces. Dit wordt bevestigd door onafhankelijke experts.

Getalsmatige feedbackmodule

Het handelingsproces kan ook in cijfers worden teruggekoppeld door de berekening van parameters die relevant zijn voor het beeldinterpretatieproces. Voorbeelden van dergelijke parameters zijn het aantal keer wisselen van de kijkrichting, het aantal keer aanpassen van het contrast of de mate van het gebruik van diverse scrolltechnieken. Door de zoekprocessen van goedscorende kandidaten te analyseren kan aan een minder goed presterende kandidaat worden teruggekoppeld in hoeverre zijn of haar zoekgedrag (op basis van de parameters) verschilt van kandidaten die dit wel goed doen.

Soorten fouten

In VQuest is het mogelijk om verschillende vraagtypen te gebruiken, bijvoorbeeld de aanwijsvraag, de kenmerkvrage (meerkeuzevraag over kenmerken van afwijkingen) en de diagnosevraag. Indien deze vraagtypen toegepast worden, al dan niet bij dezelfde casus, kan onderscheid gemaakt worden tussen de verschillende componenten van beeldinterpretatie. Op deze manier kan een inschatting gemaakt worden van de soorten fouten die de kandidaat in de toets heeft gemaakt; is er sprake van een detectiefout, een analysefout of een interpretatiefout? Indien deze vraagtypen gebruikt worden bij één casus, dan wordt het beeldinterpretatieproces als het ware nagebootst en krijgt de docent inzicht in waar in het proces het mis gaat. Deze opbouw van vraagtypen is in het huidige project niet toegepast, maar is wel al toegepast bij de beeldinterpretatie toetsing van coassistenten en de ervaringen zijn positief. De eerste ervaringen zijn beschreven in het handboek 'Digitale toetsing van radiologische beeldinterpretatie (Ravesloot 2014).

Aanbevelingen voor vervolgonderzoek

Een recent onderzoek naar zoekstrategieën in radiologische volumedatasets (CT scans van de longen) liet onderscheid zien tussen twee verschillende zoekstrategieën: drillers en scanners (Drew et al. 2013). De drillers fixeerden op één gebied in de dwarsdoorsnede en scolden vaak heen en weer in de diepte, terwijl de scanners telkens de hele dwarsdoorsnede bekeken alvorens door te scrollen naar de volgende dwarsdoorsnede, waardoor zij dus niet veel heen en weer scolden. Drillen was gerelateerd aan een hogere detectiegraad van de afwijkingen (longnodules, vlekjes in de longen). De logfiles uit het programma VQuest bevatten informatie over het heen en weer scrollen door de dataset en zouden deze zoekstrategieën kunnen blootleggen. Aangezien het te verwachten is dat een kandidaat niet alleen maar drilt of alleen maar scant, zou er een maat gedefinieerd kunnen worden voor de hoeveelheid drillen, bijvoorbeeld het aantal keer wisselen van scrollrichting (in de diepte, dus de z-as) per tijdseenheid. In toekomstig onderzoek wordt beoogd om de mate van drillen van kandidaten te vergelijken met het zoekgedrag van experts. De verwachting is dat we op deze manier rijkere, specifiekere informatie kunnen geven over het hanteren van de meest efficiënte zoekstrategie.

Hoofdstuk 4. Praktische aandachtspunten implementatie digitale radiologietoetsing

4.1 Inleiding

Gedurende het SURF-project is, naast drie digitale afnames van de voortgangstoets Radiologie, een aantal grootschalige pilots op verschillende locaties uitgevoerd. Deze digitale afnames hebben veel kennis en inzicht verschaft in benodigdheden en aandachtspunten voor succesvolle afname van grootschalige landelijke digitale (voortgangs)toetsen met beeldmateriaal in het algemeen, en de implementatie van de toetsapplicatie VQuest in het bijzonder. In dit hoofdstuk wordt een aantal belangrijke praktische aandachtspunten en aanbevelingen beschreven voor implementatie van grootschalige digitale (voortgangs)toetsing met beelden op het gebied van organisatie en toetslocatie, gebaseerd op de ervaringen van de projectleden en evaluaties van de kandidaten.

4.2 Organisatie

Een landelijk afgenomen digitale voortgangstoets vraagt om goede voorbereidingen en planning tot in detail. Logistiek gezien kan een dergelijke toetsafname een grote uitdaging zijn, zeker wanneer de toets op verschillende locaties of in verschillende zalen wordt afgenomen. Het opstellen van protocollen en draaiboeken helpt bij het verhelderen van taken en verantwoordelijkheden van betrokkenen, het stroomlijnen van processen, en vergemakkelijkt na implementatie de eventuele overdracht van taken naar de staande organisatie (zie ook hoofdstuk 5b.4). Een goed organisatorisch uitgangspunt voor implementatie van een grootschalige digitale (voortgangs)toets is één draaiboek op hoofdlijnen dat start bij de voorbereidingen van een toets (bijv. toetsdatum en locatie vaststellen) en eindigt bij het finaliseren van de toetsuitslag. Het organisatorisch draaiboek, zoals gebruikt tijdens het SURF-project, is beschikbaar via de [SURF-projectwebsite](http://www.surf.nl) (www.surf.nl), zie de link in de referentielijst. Het opstellen van een organisatorisch draaiboek is een dynamisch proces; door de toenemende ervaring van de organisatieleden en mogelijke veranderende omstandigheden zal het voor iedere toetsafname nodig zijn om de volledigheid en de actualiteit van het draaiboek te controleren en daar waar nodig aanpassingen door te voeren. Indien relevant, zal in het organisatorisch draaiboek verwezen moeten worden naar andere aanvullende draaiboeken, protocollen en instructies op detailniveau, waaronder een toetsafnameprotocol, een surveillantendraaiboek en instructie en een nood- en calamiteitenprotocol.

Toetsafnameprotocol

Een toetsafname protocol is een gedetailleerd protocol specifiek voor de toetsafnamedag. Hierin worden alle taken en verantwoordelijkheden van betrokkenen tijdens de toetsdag per tijdstip gespecificeerd. Het opstellen van een dergelijk protocol is, juist bij toetsing van grote groepen, van belang zodat de afname efficiënt en binnen de vastgestelde tijd kan verlopen. Een toetsafnameprotocol kan toets- of locatiespecifiek zijn, aangezien iedere locatie, applicatie of toets specifieke voorbereidingen met zich mee kan brengen. Een voorbeeld van een toetsafnameprotocol, zoals gebruikt tijdens het SURF-project is beschikbaar via de [SURF-projectwebsite](http://www.surf.nl) (www.surf.nl).

Surveillantendraaiboek en -instructie

De meeste opleidingen beschikken voor de afname van toetsing (zowel schriftelijk als digitaal) over een generiek surveillantprotocol. Op detailniveau kunnen de taken en verantwoordelijkheden van surveillanten voor digitale toetsing echter locatie-, toets- en applicatie-specifiek zijn. Om de toetsafname zo efficiënt mogelijk te laten verlopen, is het van belang om surveillanten tijdig op de hoogte te brengen van deze specifieke taken en verantwoordelijkheden. Afhankelijk van de toetsomstandigheden kan het opstellen van een gedetailleerd surveillantendraaiboek en instructie, die voorafgaand aan de toetsdag door de surveillanten doorgenomen dient te worden, hierbij van belang zijn. Een voorbeeld van een surveillantendraaiboek en -instructie, zoals gebruikt tijdens de eerste digitale voortgangstoets in het AMC kunt u vinden op de [SURF-projectwebsite](http://www.surf.nl) (www.surf.nl).

Nood- en calamiteitenprotocol

Een nood- en calamiteitenprotocol beschrijft hoe te handelen in geval van incidenten en calamiteiten voorafgaand, tijdens en na een digitale toetsafname, zodat:

- aantoonbaar adequaat geanticipeerd en gereageerd kan worden op incidenten en/of calamiteiten
- eenduidigheid bestaat over rollen en verantwoordelijkheden tijdens incidenten en/of calamiteiten
- onrust en onzekerheid bij betrokkenen over technische problemen, zowel voorafgaand als tijdens digitale afname, tot een minimum beperkt worden.
- interne en externe imagoschade als gevolg van incidenten en/of calamiteiten tot een minimum wordt beperkt

Een nood- en calamiteitenprotocol is bij voorkeur statisch en generiek, en moet tot in detail bekend zijn bij *alle* betrokkenen. Een voorbeeld van een nood- en calamiteitenprotocol, zoals ter advies voorgelegd aan de Examencommissie NVvR tijdens het SURF-project is te vinden op de [SURF-projectwebsite](http://www.surf.nl) (www.surf.nl).

4.3 Toetslocatie

Verskillende aspecten spelen een rol bij de keuze voor een locatie voor de afname van een grootschalige (landelijke) digitale toets. Welke aspecten uiteindelijk de toetslocatie(s) bepalen, kan per opleiding, instelling of toets verschillen. De belangrijkste aspecten en aandachtspunten die tijdens het SURF-project een rol hebben gespeeld in de keuze voor de toetslocatie, komen in Tabel 4a aan bod en worden vervolgens toegelicht.

Tabel 4a. Aspecten keuze toetslocatie

<i>Factor</i>	<i>Aandachtspunten</i>
Inrichting van de zaal (4.3.1.)	Aantal werkplekken
	Geschiktheid werkplekken
	Geluidsoverlast
	Klimaatbeheersing
	Lichtinval en verlichting
	Overzicht (surveillanten)
	Fraudepreventie
	Garderobe
	Loopruimte
Randvoorwaarden infrastructuur (4.3.2.)	Hardware eisen
	Beveiliging
	Inlogaccounts
	Distributie van de toets
	Wegschrijven van antwoorden
	Back-up toetsantwoorden
Locatie(s) (4.3.3.)	1 Locatie versus meerdere locaties
	- capaciteit
	- kosten
	- strategie
Reserveringsbeleid (4.3.4.)	Reserveringstermijn
	Vorrangsregels
	Mogelijkheid tot lange termijn afspraken
Kosten (4.3.5.)	Zaalhuur
	Kosten technische voorbereidingen
	Arbeidsuren ondersteuning tijdens toetsafname

4.3.1. Inrichting van de zaal

De inrichting van de zaal moet geschikt zijn voor de digitale toetsafname. Dit houdt, naast voldoende en geschikte werkplekken (let op voldoende reserve computers!), in ieder geval in dat geluidsoverlast tot een minimum beperkt is, de temperatuur niet te laag of te hoog is en de lichtinval en verlichting voldoen aan de gestelde eisen. Goed zicht op de werkplekken is voor surveillanten zeer wenselijk, evenals getroffen maatregelen die afkijken voorkomen. Daarnaast zijn een geschikte ruimte voor het achterlaten van jassen en tassen van kandidaten en voldoende loopruimte voor kandidaten en surveillanten een grote pré.

Inrichting van de zaal – ervaringen SURF-Project

Tijdens het SURF-project is ervaring opgedaan met een grootschalige computerzaal die specifiek voor digitale toetsing is ingericht (de digiTenT van de VU) en met kleinschalige computerzalen op het UMCU en het AMC die voor verschillende doeleinden zijn ingericht, waaronder zelfstudie en computerpractica. Alle locaties hebben specifieke aandachtspunten op het gebied van zaalinrichting. Deze aandachtspunten, zoals blijkt uit de kwalitatieve resultaten van de enquête onder arts-assistenten en bevestigd door het projectteam, zijn onder andere

AMC – 8 zalen

- Te warm in enkele zalen (>27 graden Celsius)
- Krappe werkplekken in enkele zalen
- Geluidsoverlast op de gang

UMCU – 5 zalen

- Lichtregulatie in enkele zalen
- Fraudegevoeligheid enkele zalen door het ontbreken van schotten
- Krukjes i.p.v. stoelen in 1 zaal

VU digiTenT – 1 zaal

- Gehorig
- Slechte vloerdemping
- Veel omgevingslicht
- Weerspiegeling op de beelden als gevolg van beeldschermfolie
- Slechte voorzieningen m.b.t. het opbergen van jassen en tassen

Lichtregulatie, omgevingslicht en de weerspiegeling op de beelden als gevolg van een speciale beeldschermfolie zijn bij digitale toetsafname in het algemeen belangrijke aandachtspunten, en bij toetsing met beelden in het bijzonder cruciaal. Door deze aspecten neemt de beeldkwaliteit op het computerscherm sterk af, wat de validiteit van toetsscores negatief kan beïnvloeden.

Opvallend genoeg, wordt ondanks de “dedicated” inrichting van de VU digiTenT voor digitale toetsing, de geschiktheid van de toetszaal door de arts-assistenten slechts marginaal beter beoordeeld dan de computerzalen op het AMC en UMCU (gemiddelde respectievelijk 3,5 en 3,2 op een 5 puntsschaal van zeer oneens tot zeer eens).

(Voor uitgebreide beschrijving van de evaluatieresultaten wordt verwezen naar hoofdstuk 5a.4)

4.3.2. Randvoorwaarden ICT-infrastructuur

De toetslocatie dient te voldoen aan de gestelde eisen aan de ICT-infrastructuur. Deze eisen kunnen applicatie en opleiding- (of instellings)specifiek zijn. Het gaat hierbij onder andere om eisen die gesteld worden aan de hardware (waaronder minimale schermresolutie), de beveiliging (waaronder een beschermde afname), inlog mogelijkheden, het distribueren van de toets, het wegschrijven van antwoorden en de back-up van toetsantwoorden. Zie hoofdstuk 1 voor een volledig overzicht en specificaties van de randvoorwaarden voor de ICT-infrastructuur van de digitale voortgangstoets Radiologie, zoals opgesteld binnen het SURF-project.

Randvoorwaarden ICT-infrastructuur – ervaringen SURF-project

Hardware

De hardware van alle locaties (UMCU, AMC en VU digiTent) voldoet aan de gestelde ICT randvoorwaarden (zie hoofdstuk 1). Desondanks blijkt uit de kwantitatieve evaluatieresultaten van de aios dat de laadsnelheid en de snelheid van het wisselen tussen (beeld)vragen voor de eerste digitale voortgangstoets in het UMCU en het AMC overall hoger gewaardeerd worden dan voor de tweede digitale voortgangstoets in de digiTent (respectievelijk gemiddelde scores tussen 4.13 – 4.40 en 3.07 – 3.76 op een 5 puntschaal van onvoldoende tot goed). Mogelijke verklaringen hiervoor zijn de specificaties van de hardware, die voor de werkplekken en laptops in het AMC en UMCU beter waren dan in de VU digiTent, en zwaarte van het gebruikte beeldmateriaal (tijdens de tweede voortgangstoets is één afbeelding van 200 MB gebruikt, terwijl tijdens de eerste voortgangstoets alleen beeldmateriaal onder de 150 MB is gebruikt).

Ook de algemene weergavekwaliteit van de beelden werd voor de eerste digitale voortgangstoets beduidend beter gewaardeerd dan voor de tweede digitale voortgangstoets (respectievelijk een gemiddelde score van 3,9 en 2,7 op een 5 puntsschaal van onvoldoende tot goed). Deze lagere waardering kan onder andere veroorzaakt zijn door de afkijkfolie op beeldschermen in de VU digiTent waardoor de randen van het beeld niet goed zichtbaar waren en door de kleinere beeldschermen ten opzichte van de computerzalen in het AMC en het UMCU. (Zie hoofdstuk 5a.4 voor een uitgebreide verklaring van de slechtere waardering van de beeldkwaliteit voor de tweede voortgangstoets).

Distributie toets, wegschrijven antwoorden en beveiliging

Omdat de dataload van het beeldmateriaal (MPR) bij beeldinterpretatietoetsing groot is, is het niet mogelijk om de toets web-based af te nemen. Daarom worden de toetsvragen en het beeldmateriaal in VQuest voorafgaand aan de toets over de werkplekken gedistribueerd en worden antwoorden weggeschreven naar een per locatie te specificeren server. VQuest voorziet niet in functionaliteiten die een secure afname garanderen, waardoor de beveiliging van de toets(afname) een zaak is voor de netwerk, systeem- en/of werkplekbeheerders ter plaatse.

Deze specifieke werkwijze van VQuest, vraagt om maatwerkoplossingen per toetslocatie. Alle toetslocaties (AMC, UMCU en VU) zijn hier tegen eigen uitdagingen aangelopen. Zo bleek bijvoorbeeld het gebruik van afschermingssoftware, waar het UMCU goed ervaringen mee heeft opgedaan, in het AMC geen betrouwbare oplossing voor de toetsbeveiliging. Door alle goede voorbereidingen en inzet van de betrokkenen, waaronder verschillende testen ter plaatse, is op alle locaties het distribueren van de toets, het wegschrijven van de antwoorden en de beveiliging van de toets(afname) goed verlopen.

Inloggen

De arts-assistenten Radiologie zijn werkzaam bij verschillende ziekenhuizen verspreid over het land. Dit maakt dat de meeste arts-assistenten voor de toetslocaties als externe kandidaten zijn geclassificeerd. Dit heeft gevolgen voor toetslocaties die niet gewend zijn te werken met externe partijen. Zowel het UMCU als de VU digiTent hebben voorzieningen getroffen om externe partijen via inlogaccounts toegang te geven tot de werkplekken. Het huidige ICT beleid in het AMC laat het echter, om beveiligingsrisico's, niet zonder meer toe om inlogaccounts te generen voor externen. Ook voor andere toetslocaties kan dit een belangrijk aandachtspunt zijn.

4.3.3. Locatie(s)

Bij grootschalige (landelijk) georganiseerde toetsen zal besloten moeten worden of de toets op 1 centrale plek of op meerdere plaatsen (in het land) wordt afgenomen. Het aantal benodigde werkplekken en de beschikbare capaciteit kan hierin bepalend zijn, maar ook de kosten en/of strategische keuzes bepaald door de opleiding of instelling. Gespreide afname kan bijvoorbeeld de reistijd van de kandidaten beperken, maar kan ook extra risico's en voorbereiding met zich meebrengen door verschillen in technische infrastructuur en hardware.

Locatie(s) – ervaringen SURF-project

De eerste voortgangstoets is afgenomen in het UMCU en het AMC (in respectievelijk 5 en 8 gespreide computerzalen). Tijdens de technische voorbereidingen van de eerste voortgangstoets bleek de infrastructuur, de hardware en het ICT beleid voor de twee universitair medische centra (UMC) zodanig te verschillen dat er maatwerkoplossingen bedacht moesten worden voor het distribueren van de toets, het inregelen van een server voor het versturen van antwoorden, het creëren van een beveiligde afname omgeving en de inlogprocedure (/ aanmaken van een inlogaccount). Uiteraard moesten beide toetsomgevingen ook uitgebreid getest worden om risico's tot een minimum te beperken. Het aantal voorbereidingsuren waren door de verschillende omstandigheden in de twee UMC's aanzienlijk. Een extra logistieke last was, naast de spreiding van arts-assistenten over twee UMC's, de spreiding van de arts-assistenten over verschillende computerzalen per UMC. Niet alleen was het een hele organisatie om de arts-assistenten naar de juiste zaal te begeleiden, maar per zaal moest ook voldoende ondersteuning aanwezig zijn om de toets afname in goede banen te leiden. Aanvullende aandachtspunten waren de verschillende toetsomstandigheden in de computerlokalen (zie hoofdstuk 5a.4). Voor de tweede en derde digitale voortgangstoets is gekozen voor 1 centrale toetslocatie, de VU digiTent. Deze zaal heeft tot dusver net voldoende capaciteit om alle arts-assistenten radiologie (circa 380) te toetsen. Hiermee kon een grote efficiëntieslag behaald worden voor zowel de technische voorbereidingen als de toetsafname zelf. Kosten werden gereduceerd en (ongeveer) gelijke afnameomstandigheden voor alle arts-assistenten konden worden gecreëerd.

4.3.4. Reserveringsbeleid

Omdat iedere locatie specifieke bijzonderheden met zich mee kan brengen qua IT-beleid en ICT-infrastructuur, kan een lange termijn keuze voor een toetslocatie kostenefficiënt zijn en risico's op calamiteiten en incidenten minimaliseren. Voor het maken van een lange termijn keuze, is het van belang inzage te hebben in het reserveringsbeleid van de toetslocatie(s). Hoe ver van tevoren kan de toetslocatie worden gereserveerd? Gelden er voorrangregels voor bepaalde doelgroepen? En is er ruimte voor het maken van lange termijn afspraken?

Reserveringsbeleid – ervaringen SURF-project

Het reserveringsbeleid van de huidige toetslocatie (VU digiTent) is een aandachtspunt voor continuïteit van de digitale voortgangstoets. Voor de roostering van de werkzaamheden van de arts-assistenten is het van belang om ruim van te voren toetsdata vast te leggen. Bij voorkeur worden hier lange termijn afspraken over gemaakt. Dit is momenteel niet mogelijk. Gezien de landelijke ontwikkelingen met betrekking tot het opzetten van grootschalige toetscentra is het echter de verwachting dat binnen afzienbare tijd volwaardige alternatieven (al dan niet ter back-up) zullen ontstaan.

4.3.5. Kosten

Een zeer belangrijke factor bij de keuze voor een toetslocatie zijn uiteraard de kosten. De kosten van de digitale toetsafname (waaronder zaalhuur), zullen de baten moeten rechtvaardigen (zie ook hoofdstuk 5a.2, praktische bruikbaarheid). In het geval van een toetslocatie, moet duidelijk zijn dat het niet alleen over reserveringskosten gaat maar ook over de bijbehorende arbeidsuren die nodig zijn om de toets te kunnen afnemen op de specifieke locatie(s).

Kosten – ervaringen SURF-project

Overall vielen de kosten voor digitale afname van de tweede digitale voortgangstoets (locatie VU digiTent) noemenswaardig lager uit dan voor de eerste voortgangstoets (locatie UMCU-AMC). Dit heeft te maken met de kosten van de zaalhuur en met het aantal benodigde arbeidsuren (zowel qua voorbereiding als qua ondersteuning op de toetsdag). Ter illustratie heeft de ingebruikname van de VU digiTent ten opzichte van de computerzalen bij het AMC en het UMCU, alleen al geleid tot een reductie van 35 aanwezige projectleden/ondersteunend personeel tijdens de toetsdag (eerste digitale voortgangstoets 45 medewerkers, tweede voortgangstoets 10 medewerkers).

Hoofdstuk 5a. Evalueren van een digitale radiologietoets

5a.1 Inleiding

Om digitale toetsing van beeldinterpretatie te blijven optimaliseren is het van belang om de kwaliteit van de toetsing in kaart te brengen. In termen van kwaliteit kunnen psychometrische en edumetrische kwaliteitsaspecten worden onderscheiden. Onder psychometrische kwaliteit valt bijvoorbeeld de validiteit en betrouwbaarheid van de toets. Edumetrische kwaliteitsaspecten betreffen onder meer eerlijkheid en de praktische bruikbaarheid van de toets (Poldner et al. 2012; van der Schaaf and Stokking 2008).

5a.2 Kwaliteitsindicatoren

Om de kwaliteit van de digitale VGT in kaart te kunnen brengen dient eerst vastgesteld te worden welke kwaliteitsindicatoren nodig zijn voor de evaluatie van de kwaliteit. De volgende kwaliteitsindicatoren kunnen hierbij in ogenschouw genomen worden.

Psychometrische kwaliteit

Validiteit

De validiteit van een instrument is de eigenschap dat een meetinstrument meet wat bedoeld wordt te meten. Traditioneel (en tevens volgens de huidige COTAN² richtlijnen) wordt validiteit onderscheiden in drie hoofdvormen (COTAN, 2009)³:

- Inhoudsvaliditeit
- Begripsvaliditeit
- Criteriumvaliditeit

Op deze vormen van validiteit richten we ons in het SURF-project Digitaal toetsen een extra dimensie. Hieronder volgt een toelichting van de begrippen.

- Inhoudsvaliditeit

Bij de beoordeling van een meetinstrument wordt voor de theoretische verantwoording (inhoudsvaliditeit) van het instrument nagegaan in hoeverre de herkomst van het construct (in ons geval beeldinterpretatievaardigheid) is beschreven en de te meten construct(en) zijn gedefinieerd. Het gaat om de representativiteit en dekking (volledigheid) van de items voor het te meten construct. Een toetsmatrijs is een handig hulpmiddel om dat vast te stellen. Tevens kan expertbevraging worden ingezet.

- Begripsvaliditeit of constructvaliditeit

Bij begripsvaliditeit gaat het er om te onderzoeken wat de test nu eigenlijk meet. Meet de test het bedoelde begrip of, gedeeltelijk of voornamelijk, iets anders? Begripsvaliditeit kan onder meer worden beoordeeld met behulp van een factoranalyse. Met een dergelijke analyse is het onderzoek naar validiteit niet af. We spreken daarom ook over validiteitondersteuning en niet van validiteitbevestiging. Bij begripsvaliditeit kan een onderscheid worden gemaakt tussen convergente validiteit en divergente validiteit. Convergente validiteit betreft de samenhang tussen de resultaten van de oorspronkelijke meting met de resultaten van andere metingen die hetzelfde begrip (construct) beogen te meten. Hoe hoger de correlatie tussen de scores van beide metingen, hoe beter. Bij divergente validiteit gaat het om het omgekeerde; een lage correlatie wordt verwacht met metingen van andere begrippen.

² Commissie toetsaangelegenheden Nederland

³ COTAN (2009). *Beoordelingssysteem voor de kwaliteit van tests*. Commissie Testaangelegenheden Nederland van het Nederlands Instituut van Psychologen/NIP. Beschikbaar via www.psynip.nl.

- Criteriumvaliditeit

Bij criteriumvaliditeit onderzoekt men in hoeverre er een relatie is tussen de testscore en andere kenmerken. Dat kan retrospectief, gelijktijdig (concurrente validiteit) of predictief (predictieve validiteit). Bijvoorbeeld: Het construct beeldinterpretatie hangt theoretisch samen met het begrip ruimtelijk inzicht. Over de samenhang met andere kenmerken (begrippen) kun je verwachtingen opstellen. Het is mogelijk om voor beeldinterpretatie een zogenaamd 'nomologisch netwerk' op te stellen. Een nomologisch netwerk is een logisch opgebouwd netwerk van redeneringen en verbanden waarin het instrument een plaats heeft. Verder kunnen in een populatie studenten verschillende subpopulaties onderscheiden worden. Voor de subpopulaties kunnen we verwachtingen hebben over verschillen in beeldinterpretatievaardigheid. Voor de later uit te voeren (criterium)-validiteitondersteunende analyses moet al voor afname van de meting een aantal verwachtingen worden geformuleerd over groepen die verschillen in beeldinterpretatievaardigheid en variabelen die samenhangen met dit construct. Voor de verschillende aspecten van beeldinterpretatievaardigheid moet dus ondersteuning worden gevonden voor het idee dat ze meten wat ze beogen te meten. Bevestiging van een verwacht verband of verschil tussen groepen (door toetsing: (rang-)correlatie, variantieanalyse, t-toets, etc.) kunnen we beschouwen als ondersteuning van de criteriumvaliditeit van de VGT. Als we na de analyses moeten concluderen dat verwachte verbanden niet bestaan, of dat de verschillen tussen groepen niet zijn zoals verwacht, dan kan de (spijtige) conclusie zijn dat de operationalisatie van (een aspect van) beeldinterpretatie mislukt is.

Betrouwbaarheid en onderscheidend vermogen

De betrouwbaarheid is een maat voor intra- en interbeoordelaarsovereenstemming (consistentie) op *toetsniveau*: als dezelfde beoordelaar enige weken na een toetsafname dezelfde toetsresultaten beoordeelt, of als een collega beoordelaar de beoordeling uitvoert, mogen de scores niet teveel van elkaar verschillen. Betrouwbaarheid kan geschat worden door het berekenen van bijvoorbeeld de Cronbach's alpha.

R_{AR} - en R_{IR} -waarden zijn discriminatie-indices behorende bij een item. Zij geven aan in welke mate het item hoog- en laagscorende kandidaten onderscheidt. Men noemt de R_{ir} de item-rest correlatie; het is namelijk de correlatie tussen de itemscore en de score op de resterende items van dezelfde toets. De R_{ar} is de correlatie tussen de score op de betreffende afleider en de resterende items. Een goed discriminerend item geeft een hoge positieve R_{ir} -waarde en sterk negatieve R_{ar} -waarden. Soms worden in plaats van R_{ir} - en R_{ar} -waarden ook wel R_{it} en R_{at} -waarden berekend. Dit zijn respectievelijk de correlaties tussen het item en de hele toets, en tussen het alternatief en de hele toets. Deze indices vallen hoger uit dan de R_{ir} - en R_{ar} -waarden omdat hier de itemscore in de totaalscore van de toets wordt meegeteld, waardoor de correlatie geïnflatteerd wordt.

Moeilijkheidsgraad

De moeilijkheidsgraad is een indicatie voor de moeilijkheid van toetsitems. De p-waarde is een getal tussen 0 en 1 dat de moeilijkheidsgraad van een opgave weergeeft (0.10 is erg moeilijk, 0.90 erg makkelijk). De p-waarde wordt berekend door de gemiddelde score op een opgave te delen door de maximaal haalbare score op die opgave. De p-waarde van een item is gerelateerd aan het onderscheidend vermogen. De a-waarde is de proportie kandidaten die, of het percentage kandidaten dat, bij een meerkeuze vraag de desbetreffende afleider als antwoord koos.

Edumetrische kwaliteit

Eerlijkheid

De eerlijkheid van de toets omvat onder andere of de toetsomstandigheden van de toets voor iedereen gelijk is en of het niveau van de toets in verhouding staat tot het niveau dat van de kandidaten verwacht kan worden. Er zijn verschillende manieren om de eerlijkheid van de toets te meten, waaronder het uitvoeren van gebruikersevaluaties.

Praktische bruikbaarheid

Naast de inhoudelijke aspecten is de praktische haalbaarheid van een toets een belangrijk kwaliteitscriterium. Indien er veel praktische bezwaren aan een toets kleven kan dit de toetsafname belemmeren. Een toetsprogramma dat niet intuïtief is of slechte toetsomstandigheden kunnen de acceptatie van de toets negatief beïnvloeden. In het geval van digitale toetsing met beelden is het extra van belang dat de toetsomstandigheden zo optimaal mogelijk zijn, zodat de kandidaten de beelden zo goed mogelijk kunnen beoordelen.

Naast de kandidaten zijn bij een digitale toets ook veel andere mensen betrokken: docenten, surveillanten, technici (technisch beheerders van de computers/computerzaal, beheerders van het toetsprogramma) en coördinatoren/organisatoren. Voor de deze betrokkenen moeten de voorbereidingen, kosten en toetsafname praktisch haalbaar zijn en in verhouding staan tot het belang van de toetsafname.

5a.3 Meetinstrumenten

Psychometrische kwaliteit

Item- en toetsanalyse plan

Zodra vastgesteld is welke kwaliteitsindicatoren van belang zijn, moeten methoden vastgesteld worden die nodig zijn om deze indicatoren te meten en wat de minimumeisen zijn. Een toets- en itemanalyseplan geeft inzicht in welke indicatoren van belang zijn, hoe deze gemeten dienen te worden en welke eisen daaraan gesteld worden.

Edumetrische kwaliteit

Gebruikersevaluaties

Voor het meten van de praktische bruikbaarheid van de toets kan een enquête worden afgenomen onder de kandidaten. Voor digitaal toetsen met beelden moet rekening gehouden worden met specifieke toetsomstandigheden en voorwaarden voor het gebruikersgemak van de computer en het toetsprogramma. De toetsomstandigheden moeten zo ingericht zijn dat de beelden voor iedereen goed zichtbaar zijn en idealiter dat er geen verschil is in beeldkwaliteit tussen verschillende computers en/of werkplekken. De beelden moeten in het toetsprogramma in het juiste formaat worden weergegeven. De snelheid van het laden van de beelden, het scrollen door de beelden en het wisselen tussen contrastvensters of scrollrichting moet acceptabel zijn. Deze aspecten dienen dan ook geëvalueerd te worden in de gebruikersevaluatie. De gebruikersevaluatie die is ontwikkeld om de praktische bruikbaarheid van de digitale voortgangstoets radiologie te meten staat op de [SURF-projectwebsite](http://www.surf.nl) (www.surf.nl). Voor alle andere betrokkenen bij de toets moeten de voorbereidingen en toetsafname praktisch haalbaar zijn. Om dit te evalueren zou ook een enquête afgenomen kunnen worden. Een andere mogelijkheid is om een evaluatiebijeenkomst te houden. In dit project werd na de toetsafname een evaluatieformulier ingevuld door de betrokkenen. Op de [SURF-projectwebsite](http://www.surf.nl) kunt u het gebruikte evaluatieformulier vinden (www.surf.nl). In een evaluatiebijeenkomst werden de belangrijkste verbeterpunten doorgenomen en werden oplossingen bedacht voor eventuele problemen.

5a.4 Resultaten eerste en tweede landelijke digitale voortgangstoets Radiologie

Psychometrische kwaliteit

Validiteit

- Inhoudsvaliditeit

Voor het waarborgen van de inhoudsvaliditeit van de toets zijn de eerste twee stappen uit het evidence-centered design van Mislevy (Mislevy et al. 2004) uitgevoerd. In het studentmodel is het begrip

‘beeldinterpretatievaardigheid’ gedefinieerd en theoretisch verantwoord. Op grond van het student-model en de leerdoelen uit het HORA-opleidingsplan is een taakmodel geconstrueerd met een bijbehorende toetsmatrijs. De toetsmatrijs waarborgt de representativiteit en dekking (volledigheid) van de items voor het te meten construct. Het studentmodel en taakmodel zijn beschreven in Hoofdstuk 2.

- Begripsvaliditeit of constructvaliditeit

Het is te verwachten dat een authentiekere toets het construct van beeldinterpretatie in de praktijk beter benadert en beter in staat is dit construct ook werkelijk te toetsen. De door de kandidaten gepercipieerde authenticiteit van de toets kan dan ook gezien worden als een indicatie voor begripsvaliditeit. Uit een enquête bleek dat de aios de digitale toetsen beter vinden aansluiten bij de klinische praktijk dan de voorgaande schriftelijke versies. Dit komt zowel uit de kwantitatieve als de kwalitatieve analyses. Na de eerste toets (waar nog maar elf MPR-beeldvragen in waren opgenomen bij wijze van pilot) werd door veel aios aangegeven dat ze graag meer MPR-beeldvragen in de toets wilden. In de tweede toets zaten 21 beeldvragen en werd deze opmerking niet meer gemaakt.

- Criteriumvaliditeit

De scores op de MPR-beeldvragen, 2D-beeldvragen en tekstvragen van beide digitale voortgangstoetsen stijgen naarmate de kandidaten verder gevorderd zijn in hun opleiding. In de eerste digitale VGT zien we een patroon dat vergelijkbaar is met de voorgaande schriftelijke toetsen, waarbij de scores in de eerste vier jaar van de opleiding stijgen en in het laatste jaar van de opleiding dalen of gelijk blijven. In de tweede digitale VGT zien we dat de scores op de beeldvragen een steiler verloop hebben dan de tekstvragen. Dit in tegenstelling tot de voorgaande schriftelijke voortgangstoetsen, waar de beeldvragen juist een meer afvlakkende curve toonde in vergelijking met de tekstvragen (Ravesloot et al. 2012).

Betrouwbaarheid en onderscheidend vermogen

De betrouwbaarheid van beide digitale voortgangstoetsen was hoog (Cronbach's alpha rond .90) en vergelijkbaar met de betrouwbaarheid van de voorgaande schriftelijke voortgangstoetsen. Echter, de MPR-beeldvragen hebben hier geen groot aandeel in gehad, aangezien het grootste deel van de toetsen nog uit tekstvragen en 2D-beeldvragen bestond. Uit de subanalyses blijkt dat bij beide toetsen -na correctie met behulp van de Spearman Brown formule voor het aantal vragen- de betrouwbaarheid van de MPR-beeldvragen iets hoger werd geschat dan die van de 2D-beeldvragen. Dit komt overeen met bevindingen uit het SURF-project 'Toetsing in beeld' waarin 2D- en MPR-beeldvragen werden vergeleken in de basisopleiding geneeskunde. In de huidige opzet zijn de 2D- en MPR-beeldvragen echter niet parallel geweest en het verschil in onderscheidend vermogen kan dus niet direct toegeschreven worden aan het verschil in dimensie. De geschatte betrouwbaarheid van de beeldvragen was in beide toetsen duidelijk hoger ten opzichte van de tekstvragen. Het onderscheidend vermogen van de afzonderlijke toetsitems worden niet in detail weergegeven in dit document. De vragen met een lage of negatieve Rir-waarde zijn door de examencommissie opnieuw kritisch beoordeeld en indien nodig verwijderd.

Moeilijkheidsgraad

In beide digitale voortgangstoetsen werd er hoger gescoord op de MPR-beeldvragen dan op de 2D-beeldvragen. Dit komt niet overeen met de resultaten bij de studenten uit het SURF-project 'Toetsing in Beeld': zij scoorden juist hoger op de 2D-beeldvragen. Mogelijk heeft dit te maken met de verder ontwikkelde beeldmanipulatievaardigheden van de aios, aangezien zij dagelijks met MPR-beelden werken en gewend zijn hier in meerdere richtingen doorheen te scrollen, terwijl de studenten slechts hebben geoefend met enkele MPR-beelden en verder veelal onderwezen worden met behulp van 2D-beelden. In de eerste digitale voortgangstoets werden de beeldvragen beter gemaakt dan de tekstvragen. Dit was in de tweede toets nagenoeg gelijk.

Eerlijkheid en praktische bruikbaarheid

- Toetsomstandigheden en VQuest

Het toetsprogramma VQuest werd in het algemeen goed gewaardeerd. De scores op de verschillende items over de snelheid en de gebruiksvriendelijkheid van het toetsprogramma lagen gemiddeld rond de 3 tot 4 op een schaal van 1 tot 5 (onvoldoende tot goed). De aios gaven ook aan dat ze vaker met het toetsprogramma VQuest getoetst wilden worden. Er zijn meerdere bruikbare suggesties gedaan voor verbeteringen van het toetsprogramma. Veel suggesties zijn overgenomen en reeds aangepast in het toetsprogramma.

De aios hadden verschillende klachten over de toetsomstandigheden. De opmerkingen hadden betrekking op verschillende fronten: de werkplek (te kleine tafel, oncomfortabele stoel, last van wiebelende vloer), geluidsoverlast, lichtinval (te licht), temperatuur (zowel te warm als te koud). Voor zover de situatie dit toeliet, zijn er oplossingen gevonden voor de genoemde problemen. Aan de problemen die aan de DigiTenT gerelateerd zijn wordt gewerkt om deze tot een minimum te beperken.

De eerste toets vond plaats in verschillende zalen in het UMCU en het AMC en de tweede toets in 1 grote toetszaal aan de VU. In de eerste situatie waren de toetsomstandigheden verschillend, wat nadelig is voor de eerlijkheid.

- Weergavekwaliteit beelden

De weergavekwaliteit van de beelden van de digitale toets werd voor de eerste digitale VGT beter gewaardeerd in vergelijking met de voorgaande schriftelijke toetsen. Echter, de beeldkwaliteit van de tweede VGT werd vrij negatief beoordeeld. Dit is natuurlijk een belangrijk aandachtspunt. Er zijn verschillende oorzaken voor aan te wijzen:

- 1) In de tweede versie van VQuest was het mogelijk gemaakt om meerdere series van 1 scan naast elkaar weer te geven (maximaal 4), terwijl inzoomen niet mogelijk was, doordat deze beeldbewerkingfunctie uit was gezet.
- 2) de schermen van de digiTenT hebben afkijkfolie, dat ervoor zorgde dat de randen van het beeld niet goed zichtbaar waren.
- 3) de schermen van de VU digiTenT computers zijn kleiner dan in het AMC en UMCU.
- 4) het licht was zo veel mogelijk gedimd, maar het omgevingslicht was niet optimaal (deze laatste opmerking was overigens ook een kritiekpunt bij enkele zalen in het AMC en het UMCU).

Het eerste punt is aangepakt door een 'dubbelklik' functie in te bouwen waarmee de beelden full screen zichtbaar kunnen worden gemaakt. Punt 2 en 3 zijn niet beïnvloedbaar, tenzij we uitwijken naar een andere zaal. Ten aanzien van het laatste punt is tijdens de laatste toets gebruik gemaakt van alternatieve verlichting.

- Vraagtypen

In de eerste digitale VGT werd nog geen gebruik gemaakt van andere vraagtypen. Alleen de gebruikelijke juist/onjuist/weet-niet vragen zaten in de toets. In de open opmerkingen van de enquête werd door meerdere aios verzocht om andere vraagtypen. In de tweede digitale VGT werd in de beeldvragen gevarieerd met vraagtypen: aanwijsvragen, multiple choice en meer-uit-meer vragen. In de tweede enquête werd gevraagd welke vraagtypen men geschikt vond voor het toetsen van beeldinterpretatie. Alle vraagtypen werden redelijk geschikt gevonden met een gemiddelde score van 3,3 tot 3,6 op een schaal van 1 tot 5.

- Gebruikersevaluatie andere betrokkenen

De uitwijk naar de digiTenT had een zeer positieve uitwerking op de praktische uitvoerbaarheid van de toets. De voorbereidingen en afname waren veel minder arbeidsintensief, het aantal benodigde surveillanten is drastisch afgenomen evenals de begrootte kosten.

Hoofdstuk 5b. Verduurzaming van een digitaliseringsproject

5b.1 Inleiding

Voor de verduurzaming van een (toets)digitaliseringproject is het van belang de resultaten goed te borgen. Om dit te bereiken is een aantal voorwaarden van belang:

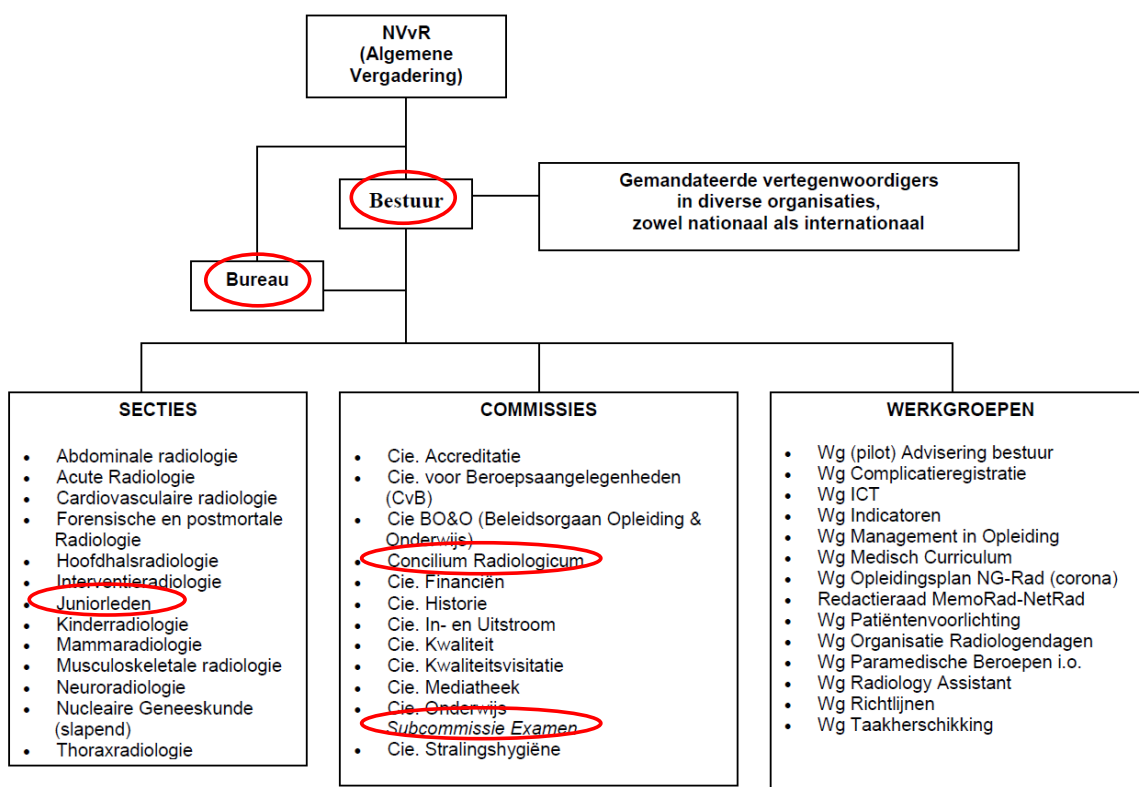
- 1) De betrokkenen in de staande organisatie en hun verantwoordelijkheden moeten duidelijk geformuleerd zijn
- 2) De processen en taakverdeling dienen duidelijk beschreven te worden en op een verantwoorde en realistische manier overgedragen te worden aan de staande organisatie
- 3) De kosteninvestering moet duidelijk zijn en acceptabel zijn voor de staande organisatie

In dit hoofdstuk gaan we in op bovengenoemde voorwaarden voor een goede borging van toetsdigitalisering in de staande organisatie, geïllustreerd met voorbeelden uit het project.

5b.2 Formulering betrokkenen en verantwoordelijkheden in de staande organisatie

Geledingen

In Figuur 5a is het organigram van de staande organisatie van dit project weergegeven, de Nederlandse Vereniging voor Radiologie (NVvR). De geledingen die betrokken zijn bij het project zijn rood omcirkeld.



Figuur 5a. Betrokken geledingen NVvR

Profielen en functieomschrijving betrokken personen

Medische beroepsverenigingen zijn vaak in grote mate afhankelijk van vrijwilligerswerk. Artsen hebben naast hun baan zitting in commissies of werkgroepen of bekleden een bestuursfunctie. Buiten een enkele reiskostenvergoeding is er geen noemenswaardige financiële compensatie voor de geleverde diensten. Dit werkt in de hand dat er een snelle turn-over is van de leden van de verschillende geledingen. De personen die de nieuwe organisatie zullen vormen, zullen dan ook veelvuldig wisselen. Het is dan ook van belang om goed vast te leggen wat het profiel en de functieomschrijving is van de betrokkenen in de staande organisatie. Op grond hiervan kan bij een eventuele wisseling een nieuwe kandidaat worden

aangewezen. Voor een geslaagde implementatie en uitvoering van de digitale VGT radiologie zijn de volgende functies essentieel gebleken: een toetscoördinator, het bureau (van de NVvR), het VQuest team, de gegevensbeheerder, adviseurs van het expertisecentrum, een contactpersoon en technische ondersteuners van de digiTent en een aanspreekpunt uit het projectteam. In Tabel 5a zijn de profielen en functieomschrijvingen van de betrokkenen in detail weergegeven.

Tabel 5a. Overzicht van profielen en functieomschrijvingen betrokkenen

a. Coördinator <u>Profiel:</u> De coördinator is lid van de examencommissie van de NVvR en heeft een stevige positie in de commissie (bijvoorbeeld als voorzitter of secretaris). Bij voorkeur is de coördinator werkend in Utrecht of Amsterdam ivm de korte lijnen met de projectleden en de VQuest ontwikkelgroep, maar dit is geen vereiste. De coördinator heeft expertise op het terrein van toetsing en heeft speciale interesse in / is bereid tot het verbeteren van de kwaliteit van de digitale voortgangstoets en de hiermee samenhangende processen. <u>Functieomschrijving:</u> De coördinator fungeert als contactpersoon tussen de examencommissie enerzijds en het bureau, het aanspreekpunt van het projectteam en de VQuest ontwikkelgroep anderzijds. De coördinator is verantwoordelijk voor de uiteindelijke toetssamenstelling en het testen van de toetspresentatie op locatie.
b. Bureau (in de persoon van de verantwoordelijke voor de VGT) <u>Profiel:</u> Heeft organisatorische en ondersteunende kwaliteiten en is in staat processen nauwgezet te bewaken. <u>Functieomschrijving:</u> Voert de NVvR protocollen uit (o.a. protocol VGT inschrijvingen verwerken, verwerken uitslagen). Bewaakt het draaiboek en treedt proactief op bij het naderen van deadlines (door signalering aan de coördinator van de examencommissie). Ondersteunt bij het uitvoeren van het draaiboek. Reserveert de toetszalen, via de contactpersoon locatie VU zalen. Verzorgt de communicatie op NetRad en communicatie met de aios.
c. Aanspreekpunt projectteam <u>Profiel:</u> Dit is een projectlid dat tijdens de looptijd van het project nauw betrokken is geweest bij de organisatie van de toets en overzicht heeft over de organisatie en de betrokkenen alsook over de expertise van de verschillende betrokkenen. <u>Functieomschrijving:</u> Fungeert na de looptijd van het project als vraagbaak voor de staande organisatie. Indien nodig kan het aanspreekpunt eventuele vragen doorspelen aan een ander projectlid dat de benodigde expertise heeft.
d. VQuest teamleden <u>Profiel:</u> Oprichters / eventuele werknemers VQuest ontwikkelgroep <u>Functieomschrijving:</u> Leveren de benodigde technische ondersteuning voor het samenstellen van de toets in VQuest en het genereren van de toetsresultaten. Zorgen dat de functionaliteit van VQuest aansluit bij de benodigde functionaliteit voor een veilige, hoogwaardige toetsafname. Testen de systeemkwaliteit van de toets op verschillende niveau's.
e. Gegevensbeheerder <u>Profiel:</u> Medewerker met kennis over en ervaring in databeheer <u>Functieomschrijving:</u> Genereren en beheren van de rapportages van de toetsresultaten per aios.
f. Adviseurs expertisecentrum <u>Profiel:</u> Experts op het gebied van medische educatie en toetsing. <u>Functieomschrijving:</u> Adviseren de examencommissie over de kwaliteit van de toets op toets- en op vraagniveau.
g. Contactpersoon toetslocatie (digiTent) <u>Profiel:</u> Heeft inzage in de beschikbaarheid van de toetslocatie en is bevoegd tot het inplannen van toetsen in het reserveringssysteem. Is op de hoogte van de stand van zaken van de toetszaal omtrent beschikbaarheid en eventuele calamiteiten. <u>Functieomschrijving:</u> Communicatie met het NVvR bureau over de reservering van de toetslocatie (digiTent) en eventuele vragen.
h. Technische ondersteuners toetslocatie (digiTent) <u>Profiel:</u> Systeembeheerders toetslocatie, die toegang hebben tot het beheer van de computers en het netwerk. <u>Functieomschrijving:</u> Verantwoordelijk voor het packagen van de toets, beveiliging van de digitale toetsomgeving en verantwoordelijk voor het wegschrijven en het beheer van de antwoordenfiles.

Verantwoordelijkheden betrokkenen

Voor de overdracht naar de staande organisatie is het noodzakelijk dat het duidelijk is wie (of welk orgaan) waarvoor verantwoordelijk is. Dit zorgt enerzijds voor toewijding ten aanzien van de taken die vervuld moeten worden en anderzijds geeft het houvast in het geval van calamiteiten, omdat men aangesproken kan worden op verantwoordelijkheden. In Tabel 5b zijn de verantwoordelijkheden weergegeven, die wij essentieel achten voor uitvoering van de digitale VGT, alsook de instanties/personen die wij aan deze verantwoordelijkheden koppelen.

Tabel 5b Verantwoordelijkheden van betrokkenen bij digitale (voortgangs)toetsing

Verantwoordelijk voor	Persoon/instantie
Inhoud van de toets	Subcommissie examen
Aanleveren toetsvragen	Subcommissie examen Beeldvragen gecoördineerd door coördinator en VQuest team Tekstvragen gecoördineerd door NVvR Bureau
Presentatie van de toets	Subcommissie examen, eindcontrole coördinator
Aanmelding en administratie kandidaten	NVvR bureau i.s.m. gegevensbeheerder
Communicatie aan aios	NVvR bureau i.s.m. gegevens beheerder
Zaalreserveringen en surveillance	NVvR bureau in samenspraak met coördinator
Bewaken deadlines (samenstellen toets, aanlevering vragen)	Tekstvragen: NVvR bureau Beeldvragen: Coördinator
Functionaliteit VQuest	VQuest ontwikkelgroep
Noodprocedure	Concilium (orgaan voor kwaliteitsbewaking opleiding)
Verwerking toetsresultaten	Subcommissie examen, uitgevoerd door expertisecentrum onderwijs UMC Utrecht / Gegevensbeheerder / VQuest ontwikkelgroep

5b.3 Beschrijving en overdracht processen en taakverdeling

Om de duurzaamheid van digitalisering van een toets te garanderen moet de hele workflow, van het voorbereiden van een toets tot de afname, ondersteund worden door vaste protocollen. Deze nieuwe workflow van een nieuwe manier van toetsen moet dan ook verankerd worden in het beleid van de betrokken instellingen. Tijdens het project is daarom een uitgebreid protocol ontwikkeld, dat ervoor zorgt dat deze verankering ook werkelijk gewaarborgd kan worden. Er is gekozen voor één organisatorisch draaiboek dat ongeveer een jaar voor de toetsafname start en doorloopt t/m enkele weken na de toetsafname. Het draaiboek vergt continue afstemming met de staande organisatie en is dan ook een dynamisch document dat op het moment van schrijven nog steeds in ontwikkeling is. In het draaiboek wordt verwezen naar andere protocollen, bv. nood- en calamiteitenprotocol en het toetsafnameprotocol (gedetailleerd protocol van de toetsafname dat elke afname verandert). De draaiboeken en protocollen zijn te vinden op de [SURF-projectwebsite \(www.surf.nl\)](http://www.surf.nl).

5b.4 Duidelijke en acceptabele kosteninvestering

Indien geen extra financiering vanuit de staande organisatie is begroot voor de voortgangstoets radiologie, moet het uitgangspunt van een toets-digitaliseringsproject zijn om de kosten van de nieuwe, digitale situatie ongeveer gelijk te houden aan de kosten van de oude situatie.

Voor een digitale toetsafname is doorgaans een andere toetslocatie nodig. Het is raadzaam om rekening te houden met tijdslots, zodat de toetsafname niet in twee verschillende tijdslots valt, wat extra kosten met zich meebrengt. In dit project heeft dit geleid tot een geringe verkorting van de toetsafnametijd. Aangezien het gebruik van de MPR-beelden juist leidt tot een toename in tijdsinvestering (dit bleek zowel uit de enquêtes als uit de logfilegegevens van de toetsafname) zijn we genooddaakt om het aantal toetsvragen ook te beperken. Hierbij moet de kwaliteit van de toets natuurlijk niet in het gedrang komen, maar gezien het grote aantal toetsvragen en de hoge betrouwbaarheden is de verwachting dat de kwaliteit van de toets bij dit itemaantal voldoende gewaarborgd zal blijven. Bij de derde voortgangstoets was het aantal vragen gereduceerd tot 180 en was de Cronbach's alpha onveranderd hoog.

Daarnaast moet rekening gehouden worden met de licentiekosten. Hierover dienen heldere afspraken gemaakt te worden met de staande organisatie en deze afspraken dienen vastgelegd te worden in een licentieovereenkomst. In een dergelijke overeenkomst moeten minimaal de volgende punten worden vastgelegd:

- betrokken partijen
- te leveren modules (van VQuest)
- duur en beëindiging van de overeenkomst
- servicelevel (aantal uren support)
- standaard updates
- meerkosten voor niet reguliere updates
- meerkosten voor extra uren support
- evaluatie-afspraken
- minimaal af te nemen toetsen
- rechten van gebruikte data
- wetenschappelijk gebruik van data en logfiles

Referenties

- Azevedo, R., Lajoie, S., Desaulniers, M., Fleischer, D., and Bret, P. (1997). "RadTutor: the theoretical and empirical basis for the design of a mammography interpretation tutor", in B. du Boulay and R. Mizoguchi, (eds.), *Artificial Intelligence in Education*. Amsterdam: IOS Press.
- Beutel, J. (2000). *Handbook of medical imaging: Physics and psychophysics*, Washington: SPIE.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain*, New York: David McKay Co Inc.
- Cooper, L., Gale, A., Darker, I., Toms, A., and Saada, J. (2009). "Radiology image perception and observer performance: how does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking." *Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment*, 7263, 72630K1-12.
- Drew, T., Le-Hoa Vo, M., Olwal, A., Jacobson, F., Seltzer, S. E., and Wolfe, J. M. (2013). "Scanners and drillers: Characterizing expert visual search through volumetric images." *Journal of Vision*, 13(10), 3.
- Hegarty, M. (2010). *Components of spatial intelligence*.
- Krupinski, E. A. (1996). "Visual scanning patterns of radiologists searching mammograms." *Acad Radiol*, 3(2), 137-44.
- Kundel, H. L., and La Follette, P. S., Jr. (1972). "Visual search patterns and experience with radiological images." *Radiology*, 103(3), 523-8.
- Kundel, H. L., Nodine, C. F., and Carmody, D. (1978). "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection." *Invest Radiol*, 13(3), 175-81.
- Kundel, H. L., Nodine, C. F., Conant, E. F., and Weinstein, S. P. (2007). "Holistic component of image perception in mammogram interpretation: gaze-tracking study." *Radiology*, 242(2), 396-402.
- Lesgold, A. M., Robinson, H., Feltovich, P., Glaser, R., Klopfer, D., and Wang, Y. (1988). *Expertise in a complex skill: diagnosing x-ray pictures*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J., and Mc Lachlan, J. (2005). "Assessment of progress tests." *Med Educ*, 39(2), 221-227.
- Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2004). *A brief introduction to evidence-centered design*. University of California.
- Morita, W., Miwa, K., Kitasaka, T., Mori, K., Suenaga, Y., Iwano, S., Ikeda, M., and Ishigaki, T. (2008). "Interaction of perceptual and conceptual processing: expertise in medical image diagnosis." *International journal of human-computer studies*, 66(5), 370-390.
- Nodine, C. F., and Kundel, H. L. (1987a). *The cognitive side of visual search*, North-Holland: Elsevier Science Publishers B.V.
- Nodine, C. F., and Kundel, H. L. (1987b). "Using eye movements to study visual search and to improve tumor detection." *Radiographics*, 7(6), 1241-50.
- Nodine, C. F., Mello-Thoms, C., Kundel, H. L., and Weinstein, S. P. (2002). "Time course of perception and decision making during mammographic interpretation." *AJR Am J Roentgenol*, 179(4), 917-23.
- Poldner, E., Simons, P. R. J., Wijngaards, G., and van der Schaaf, M. F. (2012). "Quantitative Content Analysis Procedures to Analyse Students' Reflective Essays: A Methodological Review of Psychometric and Edumetric Aspects." *Educational Research Review*, 7(1), 19-37.

- Raufaste, E., Eyrolle, H., and Mariné, C. (1998). "Pertinence generation in radiological diagnosis: spreading activation and the nature of expertise." *Cognitive Science*, 22, 517-546.
- Ravesloot, C., van der Schaaf, M., Haaring, C., Kruitwagen, C., Beek, E., Ten Cate, O., and van Schaik, J. (2012). "Construct validation of progress testing to measure knowledge and visual skills in radiology." *Med Teach*, 34(12), 1047-55.
- Ravesloot, C. J. (2014) et al. "Handboek digitale toetsing van radiologische beeldinterpretatie". City: Utrecht.
- Rogers, E. (1996). *A study of visual reasoning in medical diagnosis*, Mahwah, N.J.: Lawrence Erlbaum.
- Sadler, D. R. (1989). "Formative assessment and the design of instructional systems." *Instructional Science*, 18, 119-144.
- Sadler, D. R. (1998). "Formative assessment: Revisiting the territory." *Assessment in Education: Principles, Policy and Practice*, 5(1), 77-84.
- Sadler, D. R. (2010). "Beyond feedback: Developing student capability in complex appraisal." *Assessment and Evaluation in Higher Education*, 35(5), 535-550.
- Shute, V. J. (2008). "Focus on Formative Feedback." *Review of Educational Research*, 78(1), 153-189.
- van der Gijp, A., van der Schaaf, M. F., van der Schaaf, I. C., Huige, J. C., Ravesloot, C. J., van Schaik, J. P., and Ten Cate, T. J. (2014). "Interpretation of radiological images: towards a framework of knowledge and skills." *Adv Health Sci Educ Theory Pract*.
- van der Schaaf, M. F., and Stokking, K. M. (2008). "Developing and validating a design for teacher portfolio assessment." *Assessment & Evaluation in Higher Education*, 33(3), 1-21.
- Vincken, K. L., Ravesloot, C. J., van der Gijp, A., Mol, C., and Van der Schaaf, M. (2013). "IMAGE ME: Image interpretation Assessment using loGs for Evaluation in Medical Education". City: Utrecht.
- www.surf.nl. "Projectwebsite digitaal toetsen met beelden, een extra dimensie. ." *Beschikbaar via* <http://www.surf.nl/kennis-en-innovatie/innovatieprojecten/startdatum-2012/digitaal-toetsen-met-beelden-een-extra-dimensie.html>.

Colofon

Dit handboek maakt onderdeel uit van het project 'Digitaal toetsen met beelden, een extra dimensie', gesubsidieerd door de Stichting SURF. Dit handboek is samengesteld door een samenwerking tussen de Nederlandse Vereniging voor Radiologie (NVvR, wetenschappelijke vereniging van medisch specialisten), het Academisch Medisch Centrum (AMC) te Amsterdam, het Universitair Medisch Centrum (UMC) te Utrecht en de Universiteit Utrecht (UU). In de projectgroep waren vertegenwoordigd: de afdeling Radiologie en de afdeling Onderwijs en Studentenzaken van het AMC; de afdeling Radiologie, het Image Sciences Institute en het Expertisecentrum van het UMC Utrecht; Departement Educatie, Faculteit Sociale Wetenschappen UU; en examencommissie, bestuur en bureau van de Nederlandse Vereniging voor Radiologie.

Aan dit project is door vele personen een bijdrage geleverd, waarvoor onze dank. Onze speciale dank gaat uit naar: de examencommissie van de NVvR; het bestuur van de NVvR; het NVvR bureau: in het bijzonder Ivonne Schoenmakers; de stuurgroep van het project 'Digitaal toetsen met beelden, een extra dimensie'; Lisette van Bruggen (expertisecentrum UMC Utrecht); Marieke Krispijn (bedrijfsbureau divisie beeld UMC Utrecht); Onderwijssecretariaat Radiologie: Linda Ijskes en Amanda Groenewoud; alle aios radiologie die hebben deelgenomen aan het onderzoek; medewerkers Stichting SURF, in het bijzonder Annette Peet, Lieke Rensink, Michiel van Geloven, Jenny de Werk, Anne Floor Erdman, Esther Arrindell en Brenda van der Laan; Afdeling Multimedia divisie Beeld UMC Utrecht, in het bijzonder Karin van Rijnbach en Roy Sanders.

Grafisch ontwerp, opmaak en drukwerk: Karin van Rijnbach, Multimedia divisie Beeld UMCU
Afbeelding omslag: bewerking van een afbeelding van de eerste röntgenfoto, gemaakt door Wilhelm Conrad Röntgen in 1895.

ISBN: 978-94-6228-496-8

© 2014
1^e druk

Dit project is tot stand gekomen met steun van Stichting SURF, de organisatie die ICT vernieuwingen in het hoger onderwijs en onderzoek initieert, regisseert en stimuleert door onder meer het financieren van projecten. Meer informatie over SURF is te vinden op de website (www.surf.nl).