Eindrapportage

AdaPT, ADAPTIEVE VOORTGANGSTOETSING 1 maart 2011 – 30 juni 2014

Gegevens penvoerende instelling
Universiteit Maastricht, Faculty of Health, Medicine and Life Sciences

Naam projectleider
Dr.ir. Arno Muijtjens
Cap.gr. Onderwijsontwikkeling en Onderwijsresearch
FHML, Universiteit Maastricht
Postbus 616, 6200 MD Maastricht
tel. 043 3885745
e-mail: a.muijtjens@maastrichtuniversity.nl

Inhoudsopgave

Samenvatting

Deze eindrapportage betreft de uitgevoerde activiteiten en behaalde resultaten van het project Adapt, adaptieve voortgangstoetsing (periode 1 maart 2011 – 30 juni 2014).

Het project had als doel om na te gaan op welke wijze computergestuurd adaptief toetsen (CAT) bij voortgangstoetsing (VT) toepasbaar is en bij kan dragen aan kwaliteitsverbetering van deze toetsvorm. Bij adaptief toetsen krijgt een student een reeks vragen voorgelegd die naarmate de reeks vordert steeds beter passen bij het vaardigheidsniveau van de student. Zo'n geïndividualiseerde, automatisch samengestelde toets kan op ieder gewenst moment worden afgenomen en kan door de afstemming op het niveau van de kandidaat zeer efficiënt zijn (minder vragen nodig).
Om de onderzoeksvraag te kunnen beantwoorden is in de eerste plaats een gekalibreerde itembank nodig, dat wil zeggen een verzameling voortgangsitems die dekkend zijn voor de beoogde inhoud van de toets en waarvan de psychometrische kenmerken (moeilijkheid, discriminerend vermogen) bekend zijn. Vanaf september 2005 worden jaarlijks door de samenwerkende universiteiten van Amsterdam (VU), Groningen, Leiden, Maastricht en Nijmegen vier VT's afgenomen bij alle geneeskundestudenten. Elke VT is samengesteld uit 200 meerkeuzevragen die als gestratificeerde steekproef uit de itemvoorraad zijn getrokken conform een blauwdruk waarin per subdomein is vastgelegd hoeveel vragen de toets dient te bevatten. Elke toets bestaat uit 200 nieuwe vragen, met dien verstande dat na drie jaar een vraag weer hergebruikt mag worden.
In de voorraad van de 8x4 toetsen uit de periode 2005-2012 waren vragen die inhoudelijk voldeden dus in voldoende hoeveelheden beschikbaar. Om de psychometrische kenmerken van de vragen vast te stellen dient er een kalibratieprocedure uitgevoerd te worden. Op basis van antwoorddata voor de set van vragen worden dan de itemparameters (moeilijkheid en discriminatie) voor elke vraag geschat op basis van een IRT-model. In een vroeg stadium van het project bleek dat het niet haalbaar was om speciale kalibratieafnames voor grote groepen studenten te organiseren en daarom is uitgeweken naar een alternatief: het gebruik van de antwoorddata die verkregen zijn bij eerdere afnames met de reguliere 'papieren' VT. Daarbij treden twee complicaties op: 1) het is per VT in principe telkens een andere groep studenten die de vragen beantwoordt, en 2) bij de reguliere afname wordt gewerkt met een weet-niet optie en strafpunten voor een onjuist antwoord (formula-scoring), terwijl bij adaptieve toetsing de gebruikte IRT-modellen vereisen dat er gewerkt wordt met number-right-scoring (weet-niet optie ontbreekt). Door te kiezen voor alleen de decembertoetsen (jaar na jaar een groep deelnemende studenten in dezelfde fase van het curriculum en dus hopelijk een stabiele vaardigheidsverdeling) en bij de kalibratie gebruik te maken van de scores in het 5^{e} en 6^{e} jaar (lage percentages weet-niet) werd ernaar gestreefd deze bezwaren zoveel mogelijk te omzeilen.
Niet alle items zijn geschikt voor een adaptieve toets: alleen items die voldoende gevoelig zijn voor competentieontwikkeling zijn geschikt, d.w.z. items die een duidelijk patroon van geleidelijke groei in %correct over jaargroepen laten zien na het eerste moment van behandeling van de betreffende stof. Omdat alleen voor het Maastrichtse curriculum het behandelmoment goed te achterhalen was is het detecteren van groei-items uitgevoerd aan de hand van de scores van Maastrichtse studenten. Sommige items laten wel groei zien, maar in de vorm van een abrupte verhoging van het %correct na het behandelmoment; deze worden sprong-items genoemd. De overige items worden aangeduid als rest-items. De 1518 items uit de set van 8 decembertoetsen 2005-2012 bleken te bestaan uit 366 groei-items, 244 sprong-items en 908 rest-items.
Bij het uitvoeren van de kalibratie werd gebruik gemaakt een twee-parameter IRT model en met behulp van OPLM software werden de itemparameters geschat. Het bleek dat alleen voor de masterfase een verzameling gekalibreerde items (itembank) van voldoende omvang gevonden kon worden; de itembank bestond uit 299 groei-items. Bij de voortgangstoets wordt een twee-assige blauwdruk (19 disciplines en 17 categorieën) gebruikt die voorschrijft hoeveel vragen van elke categorie en discipline deel uit moeten maken van elke toets van 200 items. In principe kan bij zo'n toets de score per categorie en per discipline bepaald worden. Voor een adaptieve voortgangstoets

dient het aantal rapportagecategorieën 3 à 6 te zijn. In verband daarmee is vastgesteld dat de score wordt gerapporteerd over vijf subdomeinen die gevormd worden door verwante categorieën uit de blauwdruk samen te voegen. Deze supercategorieën zijn Circulatie en respiratie, Stofwisseling en voortplanting, Beweging en sturing, Mens en maatschappij en Basale en toegepaste kennis. De relatief kleine itembank bleek de blauwdruk van de voortgangstoets redelijk goed te dekken en elk van de vijf subdomeinen bleek ongeveer even sterk vertegenwoordigd in de itembank.
Op basis van de gekalibreerde itembank kon een prototype adaptieve voortgangstoets ontwikkeld worden. Daarvoor is gebruik gemaakt van het toetsservicesysteem Questify van Cito. Met dit systeem zijn simulatiestudies uitgevoerd om het adaptieve algoritme optimaal af te stellen. Dat heeft geleid tot een adaptieve toets met een vaste lengte van 100 items (20 per subdomein) met een maximaal toegestaan gebruik van items ingesteld op 80%. Dat laatste houdt in dat het algoritme er voor zorgt dat elk item in maximaal 80% van de toetsen wordt afgenomen. Op basis van de simulaties is vastgesteld dat de betrouwbaarheid van de toets ongeveer 0.90 is voor elk van de drie jaargroepen in de masterfase.
Het prototype is getest op de vijf locaties van de iVTG in pilot-experimenten met masterstudenten. Daarvoor is wervings- en instructiemateriaal ontwikkeld en goedkeuring verkregen bij de ethische commissie van de NVMO. Ter bevraging van studenten in de pilot is een vragenlijst ontwikkeld waarin voor de adaptieve en de papieren toets gevraagd wordt naar onder meer de cognitieve belasting, dekking van domeinen, de informatieve waarde van de score en in hoeverre de score overeenkomt met de eigen verwachting.
Omdat de eerste fase van de pilot (november 2013) te weinig deelnemers opleverde (N=40) is in het eerste kwartaal van 2014 een tweede fase toegevoegd. Met de verbeterde werving leverde dat een additionele 160 deelnemers op zodat het totaal kwam op N=200.
Belangrijkste conclusies van de analyse van vragenlijst, tijdregistratie en scores van deze 200 deelnemers zijn de volgende. Vergeleken met de papieren toets vinden de studenten de adaptieve toets minder moeilijk, de scores informatiever en meer conform de eigen verwachting. Echter, er wordt betwijfeld of de itembank voldoende dekking biedt van het medisch domein en het ontbreken van weet-niet optie en revisiemogelijkheden (terug naar eerdere vraag) wordt betreurd. De tijd nodig voor de beantwoording van de 100 vragen was 40±10 (M±SD) minuten, wat betekent dat de adaptieve afname in dat opzicht zeer efficiënt is: voor de 200 vragen van de papieren toets krijgen de studenten 240 minuten de tijd. Vergelijking van de score op de adaptieve toets en de gemiddelde score op de vier papieren toetsen in academisch jaar 2013-2014 levert voor de totaalscore een hoge correlatie van 0.82 op; voor de subdomeinen is de correlatie zoals verwacht mag worden (20 vragen per subomein in de adaptieve toets) lager, maar nog steeds aanzienlijk: 0.48-0.64. De correlatie van de totaalscore van de adaptieve toets met die van elk van de papieren toetsen is van vergelijkbare grootte als de correlatie tussen papieren toetsen onderling (gemiddeld: 0.77 resp. 0.78). Deze bevindingen geven aan dat er ondanks de kleine omvang van de huidige itembank grote overeenstemming is tussen de kennismeting met vier papieren toetsen van 200 vragen en de adaptieve toets van 100 vragen. Tevens is gebleken dat de betrouwbaarheid van de adaptieve toets gelijk is aan die van een papieren toets met twee keer zoveel vragen. Zeer belangrijk voor het project is dat de gevonden hoge correlaties tussen score van adaptieve en papieren toetsen ondersteunend zijn voor de validiteit van de hele keten van projectactiviteiten en bijbehorende deliverables.

Met de resultaten van dit project hebben we laten zien dat adaptieve voortgangstoetsing in het geneeskundedomein mogelijk is en met goede kwaliteit. Dat neemt niet weg dat er nog knelpunten en vragen zijn die opgelost dienen te worden voordat operationalisering mogelijk is. Herkalibratie van de itembank op grond van de antwoorddata uit de pilot is nodig om na te gaan in hoeverre de itemparameters verkregen uit de historische data (formula-scoring) valide zijn. De huidige itembank is te klein en dient substantieel uitgebreid te worden (van 300 naar bij voorkeur 2400 groei-items). Adaptieve toetsing blijkt alleen mogelijk voor de masterfase.

Vragen die zich daarbij aandienen zijn: Kunnen we op basis van de huidige voorraad items de itembank uitbreiden in de richting van de gewenste 2400 items? Zijn de groei-items voldoende representatief voor het totale medische domein, met andere woorden kunnen we zonder sprong- en rest-items voldoende dekking bereiken? Kunnen we richtlijnen ontwikkelen voor de constructie van groei-items? Moeten we naar een format waarbij de adaptieve toets open toegankelijk is voor formatief gebruik en één of twee keer per jaar onder examencondities voor summatieve doeleinden wordt afgenomen? Komt er een aparte voortgangstoets voor bachelor- en masterfase?
Vervolgonderzoek en -ontwikkeling is nodig om de vragen te beantwoorden en te werken aan het oplossen van de knelpunten.

Inleiding

Opzet en doel van het project

In dit onderzoek wordt nagegaan op welke wijze computergestuurd adaptief toetsen (CAT) bij voortgangstoetsing toepasbaar is en bij kan dragen aan kwaliteitsverbetering van deze toetsvorm. Bij adaptief toetsen krijgt een student een reeks vragen voorgelegd die naarmate de reeks vordert steeds beter passen bij het vaardigheidsniveau van de student. Zo'n geïndividualiseerde, automatisch samengestelde toets kan op ieder gewenst moment worden afgenomen en kan door de afstemming op het niveau van de kandidaat zeer efficiënt zijn (minder vragen nodig). Om antwoord te geven op de onderzoeksvraag wordt een prototypische CAT procedure ontwikkeld en getest voor de interuniversitaire Voortgangstoets Geneeskunde. De verkregen resultaten betreffen het ontwikkelde prototype, data over de performance van de procedure in pilotexperimenten en het pakket van eisen dat gemoeid is met een upgrading van een papieren voortgangstoets naar een digitale CAT versie. Deze resultaten zijn van belang voor elke toepassing van voortgangstoetsing waarbij men geïnteresseerd is in kwaliteitsverbetering door digitalisering.

Belangrijkste te behalen resultaten

Het huidige onderzoek gaat na hoe de kwaliteit van voortgangstoetsen verder verbeterd kan worden met behulp van IRT en CAT en welke gevolgen dat heeft voor de summatieve, formatieve en onderwijskundige functies van de toets. Het onderzoek wordt uitgevoerd in de context van de iVTG, maar de resultaten ervan zijn van belang voor alle bestaande en toekomstige toepassingen van voortgangstoetsing in het voortgezet en hoger onderwijs waarbij men geïnteresseerd is in kwaliteitsverbetering door digitalisering.

Het beoogde resultaat van het project bestaat uit

1. Een procedure om kandidaat-CAT-items te selecteren uit een bestaande itembank
2. Een prototype gekalibreerde iVTG itembank en CAT systeem.
3. Resultaten en conclusies van een daarmee uitgevoerde experimentele afname bij een selectieve groep studenten van verschillend expertiseniveau.
4. Het oordeel van studenten in het pilot-experiment met betrekking tot bruikbaarheid en kwaliteit van iVTG-CAT.
5. Zicht op de voor- en nadelen van deze vorm van digitalisering voor de functies van de iVTG.
6. Het pakket van eisen voor de opschaling van de huidige iVTG naar een digitale versie gebaseerd op een gekalibreerde itembank.

Het uiteindelijke doel is de verbetering van de kwaliteit van onderwijs door het beschikbaar maken van computergebaseerde adaptieve voortgangstoetsprocedures die efficiënter (nauwkeurig meten met minder items), flexibeler (tijd- en plaats-onafhankelijke individuele afname) en authentieker (multimedia) toetsen mogelijk maken.

Participerende instellingen

De participerende instellingen zijn:

- Universiteit Maastricht, capaciteitsgroep Onderwijsontwikkeling en Onderwijsresearch (O&O) van de Faculty of Health, Medicine, and Life Sciences (FHML)
    - Penvoerder en formeel uitvoerder
    - Dagelijkse uitvoering WP1a, WP2a, WP2b WP4c, WP4d, WP5a, WP5b, WP6a, WP6b, WP6c en WP6d
    - Projectcoördinatie
- Cito B.V.
    - Externe experts op het gebied van CAT en IRT
    - Dagelijkse uitvoering WP1b, WP1c, WP1d, WP3, WP4a en WP4b

Opmerking:
Het project is geïnitieerd door de Wetenschappelijke Interuniversitaire Voortgangstoetscommissie (WIV) van interuniversitaire Voortgangstoets Geneeskunde (iVTG); daarom vormt de WIV ook een belangrijk klankbord voor de ontwikkelingen in het project.

Projectperiode
AdaPT overkoepelde oorspronkelijk een periode van 36 maanden, namelijk 1 maart 2011 tot en met 28 februari 2014. In verband met een uitbreiding van de pilot-afname is de projectperiode verlengd tot 30 juni 2014.

Begroting en subsidie
Totale begroting: €173.130,-
Subsidie vanuit SURF: €120.000,-

Werkpakketten

Deze rapportage betreft de totale projectperiode 1 maart 2011 tot en met 30 juni 2014. Het project is samengesteld uit de volgende werkpakketten:

- WP1 Inrichting en voorbereiding kalibratie
- WP2 Kalibratiedata verzamelen
- WP3 Kalibratie en optimalisatie
- WP4 Simulatieonderzoek & pilotafnames
- WP5 Consequenties en pakket van eisen
- WP6 Disseminatie

WP1 Inrichting en voorbereiding kalibratie

De in dit werkpakket ingerichte projectorganisatie (WP1a) heeft goed gefunctioneerd. Naast overleg van de UM projectleden is er regelmatig overleg geweest bij Cito in Arnhem of via Elluminate/Skype. De samenwerking is op een prettige en constructieve manier verlopen en heeft geleid tot belangrijke bevindingen en resultaten voor het realiseren van een adaptieve voortgangstoets. Naast de deskundige inbreng van Cito bij diverse werkpakketten is er gebruik gemaakt van het Cito-systeem Questify voor het ontwikkelen van een prototype adaptieve voortgangstoets en de toetsafname bij de pilot-experimenten. De coulance van Cito ten aanzien van de daarmee gemoeide kosten is van belang geweest voor realiseerbaarheid van de projectdoelen binnen de gestelde financiële kaders.

Voor WP1b (opstellen toetsspecificaties en inrichten voortraject) heeft Cito een notitie geschreven waarin zij haar visie geeft op de werkpakketten 1B, 1C en 1D en het te volgen werkplan (Bijlage 1.01). Daarin wordt onder meer aangegeven dat de twee-assige blauwdruk (19 disciplines en 17 categorieën) van de voortgangstoets te fijnmazig is om er rapportagcategorieën voor de adaptieve toets op te baseren. Voorgesteld is om dit terug te brengen tot 3 à 6 domeinen. In overleg met de Wetenschappelijke Interuniversitaire Voortgangstoetscommissie (WIV) van de interuniversitaire Voortgangstoets Geneeskunde (iVTG) zijn deze vastgesteld. Het betreft een indeling in 5 supercategorieën die verkregen wordt door het samenvoegen van de oorspronkelijke 17 categorieën van de iVTG tot 5 clusters (Bijlage 1.02).

Voor WP1c (selecteren van kandidaat-CAT-items uit de bestaande iVTG itembank) is gebruik gemaakt van de resultaten in het rapport 'Haalbaarheid van computergestuurd adaptief toetsen voor voortgangstoetsing geneeskunde' van Theo Eggen en Marieke van Onna (Bijlage 1.03). De belangrijke consequentie van bevindingen in dit rapport is dat voor het selecteren van geschikte items voor elk item in de bestaande itembank bepaald moet worden in welke fase van het curriculum het aan de orde komt. Voor adaptieve toetsing zijn met name items geschikt die na het behandelmoment een gestage groei van het percentage correct over jaargroepen laten zien (zogenaamde 'groei-items'). Daarnaast zijn mogelijk ook items geschikt die na het behandelmoment een sprong in percentage correct laten zien en daarna geen substantiële groei meer ('sprong-items'). De overige items ('rest-items') worden ongeschikt geacht voor adaptieve toetsing.
Alleen in Maastricht was het mogelijk om op efficiënte wijze het behandelmoment voor elk item te bepalen. Daarom werd de classificatie van items in groei-, sprong- en rest-items uitgevoerd op basis van behandelmoment in het Maastrichtse curriculum en scoredata van Maastrichtse studenten.

Tevens is door een inhoudsdeskundige voor elk item van een voortgangstoets vastgesteld of het een 'kennisvraag' of een 'toepassingsvraag' betrof in de hoop dat die kwalificatie voorspellend kon zijn voor het kenmerk 'groei-item' (Bijlage 1.04). Dat zou nuttig kunnen zijn bij het samenstellen van de itembank. Analyse van de verkregen gegevens liet echter zien dat er geen sprake was van voorspellende waarde.

Van belang voor de selectie van items zijn de bevindingen in een onderzoek naar de relatie tussen 'groei-item' en de relevantie van een item. Daarbij is gebleken dat voor hoogrelevante items het percentage groei-items aanmerkelijk hoger is dan bij laagrelevante items (odds ratio=3.4, zie Bijlage 1.05). Dat houdt in dat als de WIV, zoals de bedoeling is, in de toekomst strenger wordt ten aanzien van de relevantie van items er ook meer items van de iVTG geschikt zullen zijn voor adaptieve toetsing.

Het uitvoeren van de oorspronkelijk geplande proefafnames (Bijlage 1.06 en 1.07) voor het verkrijgen van kalibratiedata bleek te veeleisend en risicovol te zijn. Het zou betekenen dat alle vijfde en zesdejaars studenten van de vier iVTG universiteiten een extra VT van 200 items zouden moeten maken. Problematisch is met name het via credits bereiken van serieuze deelname van studenten en daarmee het verkrijgen van valide data. Na overleg in WIV en projectgroep werd vastgesteld dat dit onhaalbaar was en werd besloten om een nieuwe koers in te slaan waarbij de kalibratie gebaseerd wordt op de beschikbare historische antwoorddata in de database van Vosys (Voortgangstoets testservice systeem). Groot voordeel daarvan is dat die data onder examencondities bij grote aantallen studenten (ongeveer 7500 per afname) zijn verkregen. Nadeel is dat de data zijn verkregen met pen-en-papier-toetsen met gebruikmaking van formula-scoring ('weet-niet optie' en strafpunten voor incorrect antwoord) terwijl de beoogde adaptieve toetsing digitaal is en gebruik maakt van number-right scoring (alles beantwoorden en geen strafpunten voor incorrect antwoord). Daarnaast is een vereiste dat de data voor de kalibratie afkomstig zijn van gelijkvaardige populaties.

De nieuwe opzet werd zo gedefinieerd dat de nadelen zoveel mogelijk werden ontlopen en aan de vereisten zo goed mogelijk werd voldaan (Bijlage 1.08). Het format van de items bij de digitale afname werd zoveel mogelijk gelijk gehouden aan het format bij de pen-en-papier-afname. Om gelijkvaardige populaties zo goed mogelijk te benaderen werd gebruik gemaakt van de data van een toetsmoment in het jaar (december). Aangezien de meerkeuzevorm van de voortgangstoets gestart is in 2005 konden we in de loop van 2012 beschikken over 8 VT's, d.w.z. 1600 items waaruit na selectie van de geschikte items een bank van voldoende omvang samengesteld kon worden.

Om de relatie tussen scores verkregen onder formula-scoring resp. number-right-scoring te kunnen onderzoeken werd voorgesteld om bij de reguliere voortgangstoetsafname van de decembertoets 2012 antwoorddata te verzamelen voor beide scoringsregels. Voor deze opzet werd uiteindelijk echter geen toestemming van de examencommissies verkregen.

Werkpakket WP1c heeft een set ruwe data opgeleverd waarbij voor de decembertoetsen van 2005-2012 per item bekend is:

- Het behandelmoment van het item in het curriculum van Maastricht
- %correct per jaargroep
- %'weet-niet' per jaargroep

WP2 Kalibratiedata verzamelen

Om aan de hand van de item-data die in WP1 zijn verzameld op efficiënte wijze te komen tot classificatie van items in groei-, sprong- en restitems is een script ontworpen en toegepast (Bijlage 2.01). In dit script zijn de regels uit Bijlage 1.08 geïmplementeerd.
Toepassing van dit script heeft geleid tot de volgende classificatie van de items:
Totaal aantal items in de decembertoetsen 2005-2012 (na vervallen items[1]): 1518
Aantal groei-items: 366 (24%)
Sprong-items: 244 (16%)
Rest-items: 908 (60%)

Voor alle items van de decembertoetsen van 2005-2012 zijn voor alle studenten van de iVTG de scoregegevens per item-student (z.g. item-reward-data) uit de database van Vosys verzameld. Dat houdt in dat voor elk van de 200 items per toets de antwoorddata van ongeveer 7500 studenten verdeeld over 6 jaargroepen beschikbaar zijn.
Vervolgens zijn deze data omgezet in bestanden die geschikt zijn om gebruikt te worden als input voor de kalibratie uitgevoerd met het programma OPLM (*One Parameter Logistic Model*, zie Bijlage 1.06).

In overleg met de WIV zijn vijf rapportagecategorieën (supercategorieën van de in de papieren toets gebruikte 17 categorieën) ontwikkeld (zie Bijlage 2.03). Deze spelen een rol bij het inrichten van het prototype adaptieve toets: bij het ontwikkelde algoritme wordt rekening gehouden met het bereiken van voldoende dekking van elk van de vijf supercategorieën. Door het prototype zal een score gerapporteerd worden voor elk van deze vijf subdomeinen.

WP3 Kalibratie en optimalisatie

De kalibratie is in diverse varianten uitgevoerd (zie Bijlage 3.04): gebruikmakend van scoregegevens van alle 6 jaargroepen samen, voor bachelorfase (jaar 1-3) en masterfase (jaar 4-6) apart, voor alleen de set groei-items en voor groei-items+sprong-items, gebruik makend van IRT modellen (OPLM) en van latent class modellen (MPLUS). Voor het samenstellen van de itembank en het ontwikkelen van het prototype bleek uiteindelijk dat de meest geschikte manier was om de kalibratie te baseren op itemparameters verkregen met OPLM voor de set van 366 groei-items en voor bachelorfase en masterfase apart. In de Masterfase geeft een set van 299 van de 366 groei-items (81.7%) een redelijke fit met OPLM gebruik makend van de conditionele maximum likelihood schattingsmethode (CML). In de Bachelorfase zijn hooguit 85 (en waarschijnlijk minder) groei-items (23.2%) die passen en deze geven dan nog steeds een beduidend slechtere fit dan de set van 299 items in de Masterfase. Hier tekent zich af dat het problematisch is om adaptieve toetsing voor de bachelorfase te ontwikkelen op basis van items die gebaseerd zijn op de einddoelstellingen van het curriculum.
Met het huidige resultaat in AdaPT was het alleen mogelijk om een protoype voor adaptieve voortgangstoetsing voor de masterfase te ontwikkelen. Dit protoype is gebaseerd op de gekalibreerde itembank van 299 groei-items.

Voor de papieren voortgangstoets wordt een fijnmazige, twee-assige blauwdruk gehanteerd van 19 disciplines en 17 categorieën (zie Bijlage 3.05). De 299 items in de gekalibreerde itembank zijn

[1] Ondanks de zorgvuldige samenstelling van de papieren voortgangstoets blijken er gemiddeld ongeveer 8 van de 200 items na afname van de toets te vervallen vanwege inhoudelijke- of vormtechnische tekortkomingen.

geselecteerd uit de 1500 beschikbare items op grond van hun meet-technische eigenschappen. In hoeverre die verzameling items de blauwdruk dekte was dus de vraag. De analyse hiervan in Bijlage 3.05 laat zien dat ondanks de geringe omvang van de itembank de blauwdruk goed gedekt wordt. De itembank bevat vragen over alle disciplines en categorieën en in aantallen die de marginalen van een blauwdruk voor 100 items (de beoogde omvang van de adaptieve toets) ruimschoots afdekken, met uitzondering van één categorie (Moleculaire en cellulaire aspecten) waarvoor een gering tekort aan items in de itembank blijkt te zijn. Wat betreft de vijf subdomeinen is er grote overeenstemming tussen de percentages in de blauwdruk en de percentages in de itembank, zie onderstaande tabel.

Subdomein (supercategorie)	%items in Blauwdruk	%items in Itembank
Circulatie en respiratie	20.5	18.4
Stofwisseling en voortplanting	23.5	27.1
Beweging en sturing	21.0	20.1
Mens en maatschappij	18.5	17.1
Basale en toegepaste kennis	16.5	17.4

De verdeling van items over de subdomeinen is ongeveer uniform, waaruit af te lezen is dat elk der subdomeinen even sterk vertegenwoordigd is in de toets en dus even belangrijk/relevant geacht wordt voor de einddoelstellingen van het curriculum.

In Bijlage 3.05 is ook gekeken naar de verschillen in dekking tussen itembank en blauwdruk op meer gedetailleerd niveau. Dan blijken er wel behoorlijke verschillen te zijn waarvan het meest in het oog springend is: in de itembank zijn items over onderwerpen van basisvakken ondervertegenwoordigd en items van klinische vakken oververtegenwoordigd. In de Discussie en conclusie in Bijlage 3.05 (p. 6) wordt daar dieper op ingegaan. Belangrijkste conclusies:

- Gebleken is dat voor alle subdomeinen in gelijke mate items geschikt voor adaptieve toetsing gevonden/geconstrueerd kunnen worden
- In vervolgonderzoek kan voor de items in de itembank worden nagegaan welke multidisciplinaire kennis van belang is om ze correct te beantwoorden
- Om meer inzicht te verwerven in de inhoudelijke en vormtechnische kenmerken van items die geschikt zijn voor adaptieve toetsing en daar bijv. richtlijnen voor constructie aan te kunnen ontlenen is vervolgonderzoek nodig. Daarbij kan gebruik gemaakt worden van de 1518 items die in AdaPT zijn geclassificeerd als groei-, sprong- en restitems en waarvan 299 groei-items uiteindelijk een zeer goed functionerende itembank hebben gevormd.

Tenslotte is nagegaan hoe de verdeling van vragen in de itembank is wat betreft herkomst (universiteit) van de vragenmaker. Omdat gegevens van het curriculum van de universiteit Maastricht gebruikt zijn bij de selectie van items voor de itembank zou een oververtegenwoordiging van 'Maastrichtse' items het gevolg kunnen zijn. De resultaten van die analyse voor de items in de itembank (Bijlage 3.05, p. 1) laten zien dat er van oververtegenwoordiging geen sprake is: Maastricht, Groningen en Nijmegen hebben een ongeveer gelijk aandeel (26%, 28%, 28%) en Leiden heeft een iets kleiner aandeel (18%). Dat laatste is plausibel omdat Leiden pas in september 2006 volledig is toegetreden tot het samenwerkingsverband iVTG en daarna in de loop van de jaren de productie van items heeft opgebouwd.

Concluderend: de in WP1, WP2 en WP3 uitgevoerde procedures om een gekalibreerde itembank voor adaptieve voortgangstoetsing samen te stellen op basis van de beschikbare voorraad items en antwoorddata van de papieren voortgangstoets heeft een kleine, maar werkbare itembank opgeleverd die als basis kan dienen voor het ontwikkelen van een prototype adaptieve voortgangstoets voor de masterfase.

WP4 Simulatieonderzoek & pilotafnames

Simulatie-onderzoek en ontwikkelen prototype

Met Cito werd overeengekomen dat het door hun ontwikkelde systeem Questify zou worden ingezet voor het ontwikkelen van het prototype en de adaptieve afname. Cito heeft daarbij gezorgd dat de kosten beperkt konden blijven tot bedragen die in de budgettering van AdaPT gedragen konden worden.

Gebruik makend van de in WP3 opgeleverde gekalibreerde itembank zijn in Questify simulaties uitgevoerd ten behoeve van de 'tuning' van het prototype. Dit heeft geleid (zie Bijlage 4.01) tot een adaptieve voortgangstoets met een vaste lengte van 100 items. De toets start met een aselect gekozen item uit de itembank van 299 items. Vervolgens worden de items aangeboden die het best passen bij de vaardigheid die de student lijkt te hebben op grond van de eerder gegeven antwoorden. Gekozen worden opgaven die maximale Fisher informatie geven bij de lopende vaardigheidsschatting. Hierbij wordt echter de restrictie gehanteerd dat elk van de 5 inhoudelijk gedefinieerde subdomeinen in gelijke verhouding in de toets voorkomt. Verder wordt in het algoritme er voor gezorgd dat elk item in maximaal 80% van de toetsen wordt afgenomen.
Na elk item wordt op basis van alle gegeven antwoorden de vaardigheid van de student, met bijbehorende schattingsfout, geschat. Hiervoor wordt de gewogen maximum likelihood methode (WML) gebruikt. De geschatte vaardigheden worden bepaald op de latente schaal die vastgelegd is in de kalibratie. Ten behoeve van de rapportage mag deze schaal lineair getransformeerd worden. Gekozen is voor een schaal die gehele getallen kan rapporteren in het bereik van 50 tot 220. De raportageschaal is als volgt gedefinieerd:

$$rapportageschaal = a \times vaardigheid\ op\ kalibratieschaal + b$$

met de transformatieconstanten a=146,843 en b=95,1542. De constanten zijn zo gekozen dat de gemiddelde vaardigheid van jaar 4 de waarde 100 krijgt en het gemiddelde plus 2 standaarddeviaties van jaar 6 de waarde 200. Getallen beneden de 50 en boven de 220 worden afgekapt. 50 als ondergrens komt ongeveer overeen met het 5e percentiel in jaar 4 en het 95e percentiel in jaar 6 ligt op ongeveer 210. De rapportage vindt plaats op basis van alle gemaakte items voor de totale toetsuitslag en voor de subdomeinen op basis van de items die gemaakt worden voor elk van die subdomeinen.

Hiermee was een goed functionerend prototype verkregen op basis waarvan de hieronder beschreven pilot-experimenten met succes konden worden uitgevoerd.

Naast dit prototype is een lineaire toets van 10 items aangemaakt om op elk van de vijf iVTG locaties te kunnen proefdraaien met Questify (Bijlagen 4.02-04).

Ontwikkelen pilot-experimenten

Ter voorbereiding van de pilot-experimenten zijn de volgende stappen uitgevoerd op alle vijf locaties van de iVTG, Amsterdam, Groningen, Leiden, Maastricht en Nijmegen:

1. Reservering computerzalen en bestellen boekenbonnen (studenten werden voor hun deelname beloond met een boekenbon van €10,-).
2. Aanvraag toestemming voor de pilot gericht aan de ethische commissie van NVMO (Bijlage 4.06)

3. Sturen brief aan onderwijsdirecteur om toestemming te vragen voor de pilot (Bijlage 4.07)
4. Organiseren werving van masterstudenten voor deelname aan de pilot (Bijlagen 4.08-11)
5. Samenstellen informatiemateriaal voor de deelnemers (Bijlagen 4.12-15)

Met name stap 4 heeft aanvankelijk problemen opgeleverd. Het streven was om in november 2013 per iVTG-locatie 60 masterstudenten-deelnemers te rekruteren. Onze inspanningen leidden echter tot een teleurstellend totaal van 40 deelnemers in november. Omdat dit aantal volstrekt ontoereikend was om conclusies te trekken omtrent de kwaliteit van de adaptieve afname heeft de projectgroep besloten om een tweede pilot te organiseren in het eerste kwartaal van 2014. Om dat mogelijk te maken was verlenging van het project met enige maanden noodzakelijk en projectgroep UM, Cito en SURF stemden hiermee in.
In de tweede pilot is het aantal deelnemers verhoogd van 40 naar in totaal 200, een aantal dat betrouwbare analyse van de kwaliteit van de adaptieve toets mogelijk maakt. Deze verbetering was te danken aan een aantal factoren:

1. De werving in 2014 was beter: de teksten waren meer to-the-point en uitnodigend, er was een aanbeveling van het onderwijsmanagement (mastercoördinator) om deel te nemen aan de pilot, de timing van de werving was beter en de werving was directer (gebruik van e-mail en/of announcement in plaats van 'ergens in een nieuwsbrief').
2. De aangeboden tijden om deel te nemen waren beter afgestemd op de doelgroep.
3. De verruiming van de pilotperiode was voor een deel van de iVTG locaties noodzakelijk om de werkzaamheden voor de pilot goed te kunnen inplannen.
4. In de tweede pilot was de urgentie voor het bereiken van hogere aantallen deelnemers nog duidelijker en konden de missers bij de eerste pilot (niet-opdagen studenten, computers die midden in een sessie uitschakelden door een tijdklok) voorkomen worden.

Lessons learned:
Bovenstaande punten vormen elk voor zich onderdeel van de lessons learned, maar de belangrijkste les is dat er bij zo'n cruciaal proces de mogelijkheid moet zijn om nog een tweede fase toe te voegen als de eerste niet toereikend is.

Ontwerpen bevraging studenten

Voor de bevraging van de studenten is een vragenlijst ontwikkeld om van de deelnemende studenten te weten te komen wat hun perceptie is van een aantal aspecten van de twee vormen van voortgangstoetsing: de papieren toets zoals ze die gewend zijn en de nieuwe computergestuurde adaptieve toets. De thema's in de vragenlijst waren (zie Bijlage 4.16) :

1. Intrinsieke cognitieve belasting (vragen 1-4),
2. Extraneuze cognitieve belasting (vragen 5-8)
3. Dekking van onderwerpen (vraag 9)
4. Vraagtekenoptie - don't know option - (vragen10-11)
5. Score zoals verwacht (vraag 12)
6. Revisiemogelijkheid (vragen 13-14)
7. Score informatief (vraag 15)

Vragen 16 en 17 betroffen de mogelijkheid om vragenboekjes mee naar huis te nemen en commentaar op vragen te leveren. Deze vragen bleken achteraf irrelevant omdat sinds kort de regelgeving bij de iVTG is veranderd en vragenboekjes niet meteen mee naar huis mogen worden genomen na de toets.

De vragenlijst was tweevoudig uitgevoerd. De eerste set van vragen (Vragenlijst vooraf, zie Bijlage 4.16) betrof de ervaringen/meningen van de student betreffende de papieren voortgangstoets. De

student werd gevraagd dit deel van de vragenlijst voorafgaand aan de adaptieve toets in te vullen. Een tweede set van soortgelijke vragen (Vragenlijst achteraf, ze Bijlage 4.16), maar dan gericht op de ervaringen/meningen van de student betreffende de zojuist beantwoorde adaptieve voortgangstoets, werd na afloop van de toets beantwoord.

Uitvoeren pilot-experimenten

Zoals eerder aangegeven is het prototype ontwikkeld in het systeem Questify van Cito en bij alle pilot-experimenten is gebruik gemaakt van deze software. Questify is een SAAS (software as a service) dienst van Cito waarbij het beheer en de afname van toetsen via het Internet verloopt. Een deelnemer logt in met een inlogcode op de Questify portal, waarna automatisch een toetspakket gedownload en uitgevoerd wordt op de lokale computer. Op de lokale computer kan vervolgens alleen deze toets gemaakt worden en zijn andere functies geblokkeerd. Questify borgt verder dat toetsen hervat kunnen worden, indien onverwachts onderbroken.

Voor het starten van een pilot-afname zijn een aantal stappen doorlopen door de toets-coördinator in Maastricht:

1. Aanlevering aanmeldingslijsten met deelnemers door iVTG-locatie.
2. Invoeren deelnemers in Questify.
3. Klaarzetten test toets (voor infrastructuur) in Questify.
4. Klaarzetten van de af te nemen toets in Questify
    a. Instellen afname tijdframe.
    b. Koppelen deelnemers aan toets.
    c. Genereren van inlogcodes voor deelnemers.
5. Terugkoppeling inlogcodes aan iVTG locatie.
6. (Laten) uitvoeren van de test toets.
7. (Laten) voorbereiden van de afname computers voorafgaand aan afname sessie.
8. Monitoren van pilot afname.

De pilot-experimenten (proefafnames adaptieve toets) zijn in drie fases gerealiseerd op alle vijf locaties van de iVTG, Amsterdam, Groningen, Leiden, Maastricht, en Nijmegen:

1. Pilot met toetscommissieleden.
2. Eerste pilots met studenten in november 2013.
3. Tweede pilots met studenten in januari-maart 2014.

De pilots met toetscommissieleden waren bedoeld om voorafgaand aan de 'echte pilot' met studenten de adaptieve toets op verschillende locaties te testen. Daarnaast boden deze proefafnames de gelegenheid aan leden van VBC's (voortgangstoetsbeoordelingscommissies) om kennis te maken met het prototype en hun eerste indrukken over het instrument kenbaar te maken.
Men was tevreden over de duidelijke instructies, de gebruiksvriendelijkheid en vlekkeloze manier waarop de toets werd afgenomen zonder technische problemen. Het moeten antwoorden, geen vraagtekenoptie of escape-mogelijkheid werd door sommigen als bezwaarlijk aangemerkt. Ten aanzien van de inhoud waren er kritische opmerkingen: veel van-hetzelfde-vragen, veel korte kennisvragen, weinig vignetvragen, veel vragen over zintuigen en medicatie. Dekking van subdomeinen werd in het algemeen redelijk in orde geacht, hoewel soms de vragen binnen een domein te eenzijdig gericht werden bevonden. Daarentegen vonden de deelnemers in het algemeen dat hun score conform verwachting was en sterke en zwakkere domeinen goed overeen kwamen met hun eigen indruk.

De aspecten die de commissieleden hebben aangeroerd zijn gebruikt bij het ontwikkelen van de vragenlijst voor studenten. Daarnaast zijn op basis van de ervaringen met deze eerste pilots de instructies voor studenten verder aangescherpt (zie Bijlage 4.15).

De pilots met studenten in november 2013 verliepen goed, het enige wat niet naar wens was waren de bereikte aantallen deelnemers. Zoals hiervoor uitvoerig toegelicht is dit aanleiding geweest voor een tweede serie pilots in januari-maart 2014 met als uiteindelijk resultaat de ervaringen/meningen en antwoorddata adaptieve toets van 200 masterstudenten verdeeld over de vijf locaties van de iVTG conform onderstaande tabel:

Locatie iVTG	Leiden (LUMC)	Maastricht (UM)	Groningen (UMCG)	Nijmegen (UMCN)	Amsterdam (VUmc)	Totaal
aantal deelnemers	47	52	21	46	32	198*

* 2 deelnemers ontbreken vanwege onvolledige identificatie resp. onvolledige Questify-data

Resultaten en conclusies analyse tijdmeting

Een belangrijke vaststelling is dat op alle locaties op een enkele uitzondering na alle studenten de sessie binnen een uur hadden afgerond. Dat houdt in dat het lezen van de instructies, invullen van de vragenlijsten en het beantwoorden van de 100 vragen minder dan een uur kostte. Deze observaties van de bij de adaptieve afname aanwezige toetscoördinatoren worden bevestigd door de tijdmetingen in Questify die aangeven dat gemiddelde en standaarddeviatie van de tijd gemoeid met het beantwoorden van de 100 vragen 40±10 minuten bedroeg. Er werden wat deze tijd betreft geen significante verschillen gevonden tussen de vijf universiteiten (Bijlage 4.17).
Bij de reguliere afname van de papieren toets van 200 vragen is de maximale duur van een sessie vier uur. De resultaten in de pilot met de adaptieve toets geven aan dat er wat de sessieduur betreft substantiële winst te behalen is met de adaptieve toets. Enige voorzichtigheid is hierbij wel op zijn plaats want de pilot-afnames zijn weliswaar uitgevoerd onder (pseudo)examenconditities maar van de bereikte scores hing geen belangrijke beslissing af zoals dat bij de papieren toets wel het geval is.

Resultaten en conclusies analyse bevraging studenten

De analyses, resultaten en conclusies van de antwoorddata van de Vragenlijst vooraf (papieren toets) en de Vragenlijst achteraf (adaptieve toets) zijn uitgebreid beschreven in een rapport (Bijlage 4.18). Hier volstaan we met het geven van een samenvatting van de belangrijkste resultaten en conclusies. De clustering van vragen tot hiervoor gepresenteerde thema's (schalen) werd ondersteund door een principale componentenanalyse van de antwoorddata van de 198 studenten die aan de pilot hebben deelgenomen. In de rapportage van de resultaten is dan ook gebruik gemaakt van deze schalen.

In onderstaande tabel zijn de meningen van de studenten samengevat. De tabel toont de plus- en minpunten in de vergelijking van de adaptieve voortgangstoets versus de papieren voortgangstoets. Omdat met name de Maastrichtse deelnemers voor sommige aspecten afwijkende meningen hadden zijn die afwijkingen in de laatste kolom vermeld.

Beoordeling door studenten van alle vijf universiteiten (N=200)		Afwijkende beoordeling door Maastrichtse studenten
Adaptief vs Papier	Voor de adaptieve voortgangstoets in vergelijking met de papieren voortgangstoets geldt:	
+	is minder moeilijk	++
+	score is informatiever	
+	score komt beter overeen met verwachting van de student	++
-	dekking is minder goed	+
-	afwezigheid vraagtekenoptie leidt tot meer onzekerheid en moeilijkere toets	
-	revisiemogelijkheid (terug naar eerdere vraag) wordt gemist	

Dat de adaptieve toets als minder moeilijk wordt ervaren is in lijn met wat verwacht wordt van een toets die beter is aangepast aan het individuele niveau van een student. Dat de score die gebaseerd is op de helft van de vragen in de papieren toets toch als informatiever en beter in overeenstemming met de verwachte score wordt bevonden, is opmerkelijk gezien de geringe omvang van de huidige adaptieve itembank. Dat de dekking minder goed wordt gevonden dan bij de papieren toets is vanwege de kleine itembank niet zo verwonderlijk. Maar kennelijk is de meet-technische kwaliteit van de voor de itembank geselecteerde items van dien aard dat de resulterende scores veelzeggend zijn voor de kennis van de student. Opvallend is dat de Maastrichtse studenten meer uitgesproken positief zijn over de moeilijkheidsgraad en het overeenkomen van de score met de verwachting en daarnaast ook positief zijn over de dekking van domeinen door de adaptieve toets. Maastrichts curriculum en scores hebben een belangrijke rol gespeeld bij het selecteren van items voor de adaptieve itembank; dat zou mede deze verschillen kunnen verklaren. Bekijken we echter de verdeling van items in de itembank naar herkomst van de auteur dan blijkt dat er wat dat betreft geen sprake is van Maastricht-bias (zie Bijlage 3.05 en sectie WP3, voorlaatste alinea).
Tenslotte missen de studenten bij de adaptieve toets de vraagtekenoptie en de revisiemogelijkheid die de papieren toets wel biedt. Beide aspecten zijn echter niet te verenigen met de eisen die een adaptieve afname stelt. Ter verdediging van het ontbreken van de vraagtekenoptie zijn er twee belangrijke argumenten: a) de bij de papieren toets toegepaste formula-scoring waarbij een vraagtekenoptie wordt gehanteerd in combinatie met strafpunten voor een incorrect antwoord heeft bepaalde nadelen: gokbereidheid en invulstrategie gaan een rol spelen en dat kan leiden tot bias in de kennismeting; b) een adaptieve toets is beter afgestemd op het kennisniveau van een individuele student waardoor een weet-niet optie minder relevant is. In een traject naar operationalisatie van adaptieve toetsing dient de discussie over het ontbreken van formula scoring, revisiemogelijkheid en andere aspecten met studenten en staf gevoerd te worden zodat belanghebbenden goed geïnformeerd zijn over de voor-en nadelen van adaptieve toetsing.

Resultaten en conclusies analyse scores studenten

Om de kwaliteit van de score op de adaptieve toets te kunnen vergelijken met die van de papieren toets is de relatie tussen beide scores onderzocht voor de studenten die hebben deelgenomen aan de pilot. Deze analyses, en de daaruit voortvloeiende resultaten en conclusies zijn uitgebreid beschreven

in een rapport (Bijlage 4.19). Hier volstaan we met het geven van een samenvatting van de belangrijkste resultaten en conclusies.
De analyse was met name gericht op het vaststellen van de (Pearson product moment) correlatie tussen de adaptieve score en de corresponderende score bij de papieren voortgangstoets. Wat de adaptieve toets betreft waren zes scores beschikbaar, een totaalscore en scores voor elk van de vijf subdomeinen weergegeven in onderstaande tabel:

Omschrijving domein	Afkorting
Totaal	Tot
Basale en toegepaste kennis	Bas
Mens en maatschappij	Men
Beweging en sturing	Bew
Stofwisseling en voortplanting	Sto
Circulatie en respiratie	Cir

De zes adaptieve scores werden voor alle deelnemers geëxporteerd uit Questify. Voor deze studenten zijn de scores in de vier papieren voortgangstoetsen (september, november, februari en mei) in het academisch jaar 2013-2014 verzameld. Vervolgens werd de gemiddelde score (formula-scoring) over de vier toetsen (totaal en per subdomein) berekend en deze werd gebruikt als indicatie van het kennisniveau volgens de papieren toets.
De correlatie tussen adaptieve score en gemiddelde papieren score voor alle 198 studenten is de belangrijkste indicator. Omdat de selectie van items in de itembank is gebaseerd op groeipatronen in scores en curriculuminformatie van studenten in Maastricht is nagegaan of de correlatie voor die universiteit hoger uitvalt dan voor de andere universiteiten.
Naast de correlaties tussen adaptieve score en gemiddelde papieren score is nagegaan hoe de gemiddelde correlatie van adaptieve score met de score van elk van de vier papieren toetsen afzonderlijk zich verhoudt tot de gemiddelde correlatie tussen de vier papieren toetsen onderling. Als die correlaties van dezelfde orde van grootte zijn dan is dat een indicatie dat de adaptieve toets met de helft van de vragen het totale domein even betrouwbaar meet als een papieren toets.
Van de 200 studenten waren er 10 waarvoor de data voor deze analyse niet volledig waren, dus de analyse werd uitgevoerd voor in totaal N=190 studenten. De verdeling over de curriculumjaren (bachelor 1-3, master 4-6) was als volgt:

Jaargroep	Aantal	Percentage
3	5	2.6
4	74	38.9
5	50	26.3
6	61	32.1
Totaal	190	100.0

De belangrijkste resultaten van de analyse zijn als volgt (a: adaptieve toets; p: papieren toets; pg is gemiddelde van de 4 p-toetsen):

1. Correlatie a-pg totaal:
   0.82 ($p<0.0005$)
2. Correlatie a-pg totaal per universiteit (Lei, Maa, Gro, Nijm, Ams):
   0.82, 0.80, 0.67, 0.84, 0.85 ($p<0.0005$, $p<0.0005$, $p<0.002$, $p<0.0005$, $p<0.0005$)
3. Correlatie a-pg subdomeinen:
   0.48-0.64 (alle $p<0.0005$)
4. Correlatie a-p totaal, gemiddeld over de vier p-toetsen:
   0.77 (alle $p<0.0005$)
5. Correlatie p-p totaal, gemiddeld over de zes p-toets-combinaties:
   0.78 (alle $p<0.0005$)

Discussie en conclusies:

De hoge correlatie tussen de adaptieve totaalscore en de gemiddelde totaalscore van de vier papieren toetsen (ad 1) geeft aan dat de meting van het kennisniveau van de student gemeten met vier papieren toetsen van 200 vragen grote overeenkomsten heeft met de meting van het kennisniveau met een adaptieve toets van 100 vragen. Hoe groot de overeenkomst is wordt nog duidelijker als in aanmerking wordt genomen dat het om de correlatie gaat van twee met meetfout behepte indicatoren. Ervan uitgaand dat beide scores een betrouwbaarheid van om en nabij de 0.9 hebben (Bijlage 4.01 en Wrigley et.al[2]) leidt toepassing van de attenuation formula (zie Crocker&Algina[3]) tot een geschatte correlatie 0.9 tussen de true scores, de scores zonder meetfout.
Uit de correlaties a-pg totaal berekend per universiteit (ad 2) blijkt dat er van een relatief hoge correlatie voor Maastricht volstrekt geen sprake is. De relatief lage correlatie voor Groningen is waarschijnlijk te wijten aan een te geringe spreiding van kennisniveau (restriction of range) in de kleine Groningse deelnemersgroep. De conclusie is dat de correlatie a-pg totaal van 0.82 (ad 1) representatief geacht mag worden voor alle universiteiten.
De correlaties a-pg voor subdomeinen (ad 3) zijn een stuk lager dan die voor totaal, maar nog steeds aanzienlijk. Aangezien de subdomeinscores onderling ook hoog correleren is het plausibel dat de totaalscore (ongeveer vijf maal zoveel vragen) veel minder ruis bevat en daardoor een veel hogere correlatie a-pg laat zien dan elk van de subdomeinen.
De overeenstemming tussen de gemiddelde correlaties ad 4 en 5 zijn een indicatie dat met de 100 vragen in de adaptieve toets het kennisniveau van de studenten even betrouwbaar wordt gemeten als met de 200 vragen in de papieren toets.

[2] Wrigley, W., Van der Vleuten , C. P. M., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: *AMEE Guide No. 71. Medical Teacher, 34*, 683-697.
[3] Crocker, L., & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. Fort Worth: Harcourt. (p.237)

WP5 Consequenties en pakket van eisen

Met de resultaten in dit project hebben we laten zien dat adaptieve voortgangstoetsing in het geneeskundedomein mogelijk is en met goede kwaliteit. De goede kwaliteit blijkt uit de (voor een deel) positieve oordelen van staf en studenten over de adaptieve toets (minder moeilijk, score informatiever, score komt beter overeen met verwachting) en uit de hoge correlaties tussen adaptieve scores en papieren-toets-scores van de deelnemers aan de pilot.

Daarnaast zijn er indicaties dat de kennismeting met de adaptieve toets beduidend efficiënter is: gemiddeld werd de uit 100 vragen bestaande toets door de studenten beantwoord in 40 minuten en de resultaten van de correlatieanalyse geven aan dat de betrouwbaarheid van de score vergelijkbaar is met die van een papieren toets met 200 vragen. Als deze afnameduur indicatief is voor de duur vereist bij een 'echte' afname voor summatief gebruik dan is het mogelijk om met een batchgewijze afname in een computerpark van 250 werkstations binnen 1 à 2 dagen de adaptieve voortgangstoets bij 1500 studenten af te nemen.
Bij gebruik van de adaptieve voortgangstoets hoeven echter de afnames niet geconcentreerd te worden op één of enkele dagen en hoeft de afname ook niet tegelijkertijd voor de vijf universiteiten plaats te vinden zoals nu wel het geval is bij de papieren toets. Adaptieve toetsing biedt dus goede mogelijkheden voor het verlichten van de huidige logistieke belasting door meer flexibiliteit ten aanzien van tijd en plaats van de afname. Ook in onderwijskundig opzicht is die flexibiliteit interessant omdat het beter past bij grotere variatie in leertrajecten en meer zelfsturing van studenten.

Hoewel niet opgenomen als projectdoel was het de bedoeling van de projectgroep om de evaluatie van het prototype af te ronden met een herkalibratie van de itembank op basis van de antwoordgegevens verzameld in de pilot. Helaas is dit niet binnen de projectperiode gelukt omdat de benodigde gegevens nog niet door Questify geëxporteerd konden worden. Hoewel de gevonden hoge correlaties tussen adaptieve scores en papieren toetsscores ondersteunend zijn voor de validiteit van de itemparameters, is een herkalibratie van nut om dat nog grondiger te evalueren. De huidige itemparameters zijn verkregen door kalibratie op basis van historische onder formula-scoring condities verkregen data bij de papieren voortgangstoets. De vraag is of die parameters substantieel wijzigen bij een herkalibratie op basis van de onder number-right condities verkregen antwoorddata bij de pilot. De herkalibratie zal uitgevoerd worden na afloop van het project zodra de exportmodule van Questify functioneert.

Een belangrijke beperking is dat het alleen mogelijk was om adaptieve toetsing te ontwikkelen voor de masterfase. Kennelijk zijn de meet-technische eisen die een adaptieve toets stelt (items moeten voldoende gevoelig zijn voor competentiegroei) zodanig dat het zeer moeilijk zo niet onmogelijk is om dat in de bachelorfase te realiseren voor items die gericht zijn op de einddoelen van het zesjarig geneeskunde curriculum.

Een tweede beperking is de geringe omvang van de huidige itembank (N=299). Dat daarmee een adaptieve kennismeting is verkregen met deze goede kwaliteit is opmerkelijk. Dat resultaat geeft aan dat het selectieproces een meet-technisch adequate itembank heeft opgeleverd waarin ook nog, zonder dat daar op gestuurd is, alle vijf subdomeinen evenredig vertegenwoordigd zijn. Kennelijk is het in alle subdomeinen in gelijke mate mogelijk om items geschikt voor adaptieve toetsing te vinden/construeren. Dat neemt niet weg dat een itembank voor een adaptieve toets in een operationele setting veel omvangrijker moet zijn, eerder in de richting van 2400 items (zie Bijlage 5.01). Uitgaande van de huidige oogst van 300 geschikte items uit 8 voortgangstoetsen zouden we met de huidige voorraad items (2005-2013: 4.5x8 voortgangstoetsen) kunnen komen tot 4.5x300=1350 items. Daarmee kom je op iets meer dan de helft van de in Bijlage 5.01 voorgeschreven aantallen voor een jaar toetsen zonder overlap. Een aantal kanttekeningen dient daarbij gemaakt te worden: a) het probleem van kalibratie met data van populaties met een

verschillende vaardigheidsverdeling (over de vier toetsmomenten in een academisch jaar) dient oplosbaar te zijn; b) het selecteren van groei-items op basis van scorepatronen dient uitvoerbaar te zijn voor groepen in verschillende curricula (en liefst zonder dat informatie nodig is over het behandelmoment omdat dat een complex en arbeidsintensief proces vergt); c) het aantal identieke items (door hergebruik van items na drie jaar) in de verzameling geselecteerde items moet niet te groot zijn anders worden de 1350 items mogelijk niet gehaald.
Voor de noodzakelijke aanvulling en verversing van de itembank is het nodig dat de constructie van nieuwe items wordt gecontinueerd en dat die items ge-pre-test worden in de adaptieve afname (als items die niet meetellen voor de score). De constructie van items dient dan met name gericht te zijn op items geschikt voor een adaptieve toets in de masterfase. Daarvoor zijn nieuwe richtlijnen voor de constructie van items nodig want van de huidige geconstrueerde items blijkt slechts 20% (300 van 1518 items) geschikt te zijn voor adaptieve toetsing. Om die richtlijnen te ontwikkelen is onderzoek nodig waarbij gebruik gemaakt kan worden van de huidige verzameling van 1518 items die als groei-, sprong- en rest-items zijn geclassificeerd. In die leerset zijn mogelijk inhoudelijke en/of vormtechnische kenmerken te vinden die helpen bij het ontwikkelen van constructie-richtlijnen voor groei-items die geschikt zijn voor adaptieve toetsing.
Het vergaren van kalibratiegegevens voor nieuwe items en de onderwijskundige waarde van formatieve toetsing kunnen beide gediend worden door een opzet waarbij gedurende het jaar de adaptieve toets open toegankelijk is voor formatief gebruik en één of twee keer per jaar onder examencondities voor summatieve doeleinden wordt afgenomen. Cito is in een ander project bezig met het ontwikkelen van zo'n opzet en ziet daarvoor mogelijkheden bij de adaptieve voortgangstoets geneeskunde.

De inhoudelijke analyse van de huidige verzameling van items en met name de items in de itembank kan ook meer licht werpen op de vermeende eenzijdigheid van de items in de huidige itembank en de twijfel aan voldoende dekking. Uiteraard mag verwacht worden dat uitbreiding van de itembank, bijvoorbeeld langs de lijnen zoals hierboven aangegeven, zal leiden tot betere dekking. Maar dan kan nog steeds de volgende vraag aan de orde komen: kan in de masterfase volstaan worden met alleen de adaptieve afname of dienen daarnaast de sprong- en rest-items ook bevraagd te worden om voldoende dekking te krijgen? Dat laatste zou betekenen dat naast de adaptieve afname nog een additionele toets dient te worden afgenomen in de vorm van een bemonstering van het kennisdomein zoals dat nu ook bij de papieren toets gebruikelijk is. Uiteraard is het aantrekkelijker als de dekking wel voldoende geacht kan worden en er volstaan kan worden met alleen een adaptieve toets. Wat dat betreft is het bemoedigend dat eerder gebleken is (zie WP1) dat onder hoog-relevante items meer groei-items worden aangetroffen dan onder laag-relevante items.
Op grond van de huidige resultaten moet vastgesteld worden dat een adaptieve voortgangstoets in de bachelorfase niet mogelijk is. Verder ontwikkelen van een adaptieve voortgangstoets voor de masterfase houdt dan in dat het huidige voortgangstoets-continuüm over het hele zesjarige curriculum losgelaten wordt. Er ontstaat dan een situatie waarbij bachelorfase en masterfase elk een eigen voorgangstoets hebben. De vraag die daarbij rijst is: zou een adaptieve toets in de bachelorfase wel mogelijk zijn voor een verzameling items die niet gericht is op de einddoelen van de het zesjarig curriculum maar op de einddoelen van de driejarige bachelorfase?

De studenten (en sommige stafleden) hebben bij de bevraging tijdens de pilot aangegeven dat ze het een nadeel vinden dat bij de adaptieve toets geen mogelijkheid was om terug te komen op het antwoord op een eerdere vraag en dat de vraagtekenoptie ontbreekt. Het eerste is onontkoombaar bij een adaptieve toets, maar aan het tweede bezwaar kan tegemoet gekomen worden, door het toevoegen van een weet-niet optie en/of het in het programma mogelijk maken om een vraag onbeantwoord te laten en door te gaan naar de volgende vraag.

De huidige adaptieve toets bevatte uitsluitend vragen die qua inhoud en vormgeving identiek waren aan de vragen in de papieren toets. De reden daarvoor is simpel: bij het samenstellen van de itembank van de adaptieve toets is geput uit de voorraad vragen van de papieren toets met als bijkomend voordeel dat de twee toetsvormen goed vergeleken konden worden. Echter, een van de belangrijke voordelen van een computer-gebaseerde adaptieve toets is daardoor niet aan de orde gekomen in de beoordelingen bij de pilot: meer mogelijkheden voor gebruik van multimedia in de vragen, waardoor de vragen authentieker en gevarieerder qua vorm en inhoud kunnen zijn. Verwacht mag worden dat staf en studenten die ruimere mogelijkheden positief zullen waarderen.

De hoge correlaties die we vinden zijn belangrijk voor het vaststellen van de kwaliteit van de adaptieve kennismeting. Daarnaast, en voor het project AdaPT minstens zo belangrijk, zijn ze ondersteunend voor de validiteit van de projectactiviteiten (en daaruit voortvloeiende deliverables) die de ontwikkeling van het prototype en de uitvoering van de pilot mogelijk hebben gemaakt. Dat betreft de hele keten van procedures te beginnen bij de selectie van items voor de itembank, de kalibratie van de itembank, het afname-algoritme, de software, de afname zelf en tenslotte de export van adaptieve scores uit Questify. Als in die keten ergens belangrijke steken zouden zijn gevallen dan is het hoogst onwaarschijnlijk dat aan het eind nog een hoge correlatie tussen adaptieve toetsscore en papieren toetsscore zou worden gevonden.

De resultaten, conclusies en vragen in dit rapport zullen worden besproken in de wetenschappelijke commissie (WIV) van de iVTG en op basis daarvan zal bepaald worden wat de vervolgstappen voor dit project zijn. In dat verband is het belangrijk om te melden dat van de zijde van Cito aangegeven is dat er interesse is om de samenwerking na het project voort te zetten en dat binnen de UM en binnen de iVTG de ontwikkeling van digitale toetsing hoog op de agenda staat.

Met de activiteiten in dit project zijn belangrijke stappen gezet voor het ontwikkelen van een adaptieve voortgangstoets en zijn kennis en inzicht vergroot. Zoals hierboven is aangegeven heeft dat geleid tot nieuwe vragen die in vervolgonderzoek aan de orde moeten komen alvorens overgegaan kan worden naar een fase van operationalisering.

WP6 Disseminatie

De volgende activiteiten en producten betreffen de disseminatie van kennis, procedures en materialen die in het kader van het project zijn ontwikkeld.

Schrijven eindrapport

In verband met de verlenging van de projectperiode is met de Commissie Projectbewaking van SURF afgesproken dat de inhoudelijke eindrapportage op 1 augustus 2014 wordt opgeleverd en de financiële eindrapportage op 1 oktober 2014.
In de maand juli 2014 is de inhoudelijke eindrapportage voorbereid en afgerond.

Congresbijdragen en wetenschappelijke artikelen

In verband met dit project zijn/worden de volgende congresbijdragen/webinars door projectleden verzorgd:

1. Arno Muijtjens. Adaptieve voortgangstoetsing. Presentatie OWD[4] 2011, Utrecht. (Bijlage 6.03)
2. Jean van Berlo, Annemarie Camp, Marieke van Onna, Theo Eggen, Jeroen Donkers, Arno Muijtjens. The Netherlands Interuniversity Progress Test in Medicine – further developments. Presentation AMEE[5] 2012, Lyon, France. (Bijlage 6.04).
3. Arno Muijtjens, Marieke van Onna, Theo Eggen, Jean van Berlo, Tineke Krommenhoek, Lambert Schuwirth. Item growth patterns and item relevance in relation to adaptive progress testing. AMEE 2012, Lyon, France. (Bijlage 6.05).
4. Jean van Berlo, Marieke van Onna, Theo Eggen, Arno Muijtjens. AdaPT: Ontwikkeling van een adaptieve voortgangstoets geneeskunde. NVMO[6] 2012, Maastricht. (Bijlage 6.06)
5. Arno Muijtjens, Marieke van Onna, Theo Eggen, Jean van Berlo, Tineke Krommenhoek, Lambert Schuwirth. Item-groeipatronen en item-relevantie bij adaptieve voortgangstoetsing. NVMO 2012, Maastricht. (Bijlage 6.08).
6. Arno Muijtjens, Jean van Berlo, Jimmie Leppink, Annemarie Camp, Marieke van Onna, Theo Eggen, Jeroen Donkers. De interuniversitaire voortgangstoets geneeskunde - ervaringen en recente ontwikkelingen. OWD 2012, Rotterdam. (Bijlage 6.10).
7. Jean van Berlo, Marieke van Onna, Theo Eggen, Hetty Snellen, Jimmie Leppink, Arno Muijtjens. AdaPT: Development of an adaptive progress test in medicine. SHE Academy 2013, Maastricht. (Bijlage 6.11).
8. Arno Muijtjens. Progress Testing – concept, history, and recent developments. IAMSE[7] Web seminar 2013. (Bijlage 6.12).
9. Jean P.M. van Berlo, Jimmie Leppink, Theo J.H.M. Eggen, Arno M.M. Muijtjens. A Computerized Adaptive Test (CAT) for the Dutch Medical Curriculum: Development and Experience. CAA 2014, Zeist. (Bijlage 6.18).
10. Muijtjens, A. M. M., Leppink, J., Van Berlo, J.P.M., Meiboom, A.A., Tio, R.A., Eggen, T.J.H.M. Computerized Adaptive Progress Testing in the Medical Domain: A Study of Students' Experiences. AMEE2014, Milan, Italy. (Bijlage 6.13).

[4] Onderwijsdagen
[5] Association of Medical Education in Europe
[6] Nederlandse Vereniging voor Medisch Onderwijs
[7] International Association of Medical Science Educators

Aan de volgende wetenschappelijke artikelen/hoodstuk in boek op het gebied van de ontwikkeling van voortgangstoetsing is in de projectperiode een bijdrage geleverd door projectleden:

1. Muijtjens, A. M. M., & Wijnen, W. H. F. W. (2014). Toetsen met voortgangstoetsen. In H. Van Berkel, M. Bax & D. Joosten-ten Brinke (Eds.), Toetsen in het hoger onderwijs (3e ed.). Houten, Netherlands: Bohn Stafleu van Loghum. (Bijlage 6.15).
2. Muijtjens, A. M. M. (2013). Voortgangstoetsing. *Examens*(februari NR 1), 7-10. (Bijlage 6.16).
3. Wrigley, W., Van der Vleuten , C. P. M., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher, 34*, 683-697. (Bijlage 6.17).

Twee wetenschappelijk artikelen over resultaten en conclusies van het project AdaPT zijn in voorbereiding: een artikel over de student percepties en praktische bevindingen en een tweede artikel over de meettechnische aspecten. Naar verwachting worden deze artikelen in het najaar van 2014 ingediend.

Voortgangsrapportage voor de iVTG-leden

Op diverse vergaderingen van de Wetenschappelijke Interuniversitaire Voortgangstoetscommissie (WIV) van de iVTG is verslag uitgebracht over de vorderingen van het project en zijn bevindingen en koerswijzigingen besproken. In de pilot-periode is intensief contact geweest met de iVTG toetscoördinatoren op elk van de vijf universiteiten om te zorgen dat de pilot in goede banen werd geleid.
Dit eindrapport wordt ook rondgestuurd naar de leden van de Wetenschappelijke Interuniversitaire Voortgangstoetscommissie (WIV) van de iVTG en zal in najaar 2014 in de WIV besproken worden.

TTL activiteiten

In het kader van TTL activiteiten is vanuit AdaPT deelgenomen aan diverse SURF TTL bijeenkomsten. Met het pijler 2 TTL project VGTogether heeft regelmatig uitwisseling van informatie plaatsgevonden.

In het kader van de gezamenlijke activiteiten in TTL-verband is gewerkt aan een checklist van procedures voor het ontwikkelen van toetsitems, is in twee sessies meegewerkt aan een interview over AdaPT en is de Vragenlijst AdaPT met toelichting beschikbaar gesteld.

Wijziging Controlling Document

Naar aanleiding van de koerswijziging aangegeven in de rapportage van januari-maart 2012 is aan de projectleider verzocht om het controlling document (CD) aan te passen aan de nieuwe situatie. De nieuwe versie van het CD (d.d. 23-08-2012) is mede tot stand gekomen op basis van adviezen van de reviewer van het project, Dr. J.T. Goldschmeding, en is op 13-09-2012 door Drs. A. Peet akkoord bevonden.

Aanpassing planning

Met Drs. A. Peet is (e-mail d.d. 23 december 2013) overeengekomen dat in verband met de verlengde pilot de planning van AdaPT wordt aangepast: het project wordt verlengd tot 31 juni 2014, de inhoudelijke eindrapportage wordt op 1 augustus 2014 aangeleverd en de financiële eindrapportage wordt op 1 oktober 2014 aangeleverd.

Kennisdisseminatie

Voor een overzicht van de activiteiten op het gebied van kennisdisseminatie verwijzen we naar de producten en activiteiten vermeld in sectie WP6 onder 'Congresbijdragen en wetenschappelijke artikelen', 'Voortgangsrapportage voor de iVTG-leden' en 'TTL activiteiten'.

Deskundigheidsbevordering

In het kader van deskundigheidsbevordering hebben Arno Muijtjens (UM) en Jean van Berlo (UM) op aanraden van Cito deelgenomen aan de cursus Multidimensionele Item Response Theory (MIRT). Deze cursus heeft de nodige kennis opgeleverd voor het toepassen van zowel unidimensionele als multidimensionele IRT modellen voor computergestuurd adaptief toetsen (CAT) van de interuniversitaire voortgangstoets (iVTG). Een overzicht van de inhoud van de cursus is te vinden in Bijlage 6.01.

Van 20-6-2011 t/m 24-6-2011 heeft een studiereis plaatsgevonden naar de National Board of Medical Examiners (NBME) en The Medical Council of Canada (MCC) waarbij de NBME op 21-6-2011 werd bezocht en de MCC op 23-6-2011. NBME is een onafhankelijke instelling die zich bezig houdt met (voortgangs)toetsing op het gebied van medisch onderwijs in voornamelijk de Verenigde Staten, maar ook daarbuiten. MCC is de Canadese tegenhanger van de NBME en is zodoende een vergelijkbaar expertise centrum.

Tijdens deze studiereis is informatie verkregen over toepassingen van IRT en computergestuurd adaptief toetsen (CAT) vooral gericht op voortgangstoetsing. De discussie ging daarin over "do's" and "don't do's" op dit gebied en hoe CAT toegepast kon worden in de context van de iVTG. Een verslag van deze studiereis is te vinden in Bijlage 6.02.

Deze studiereis heeft waardevolle informatie en inzichten opgeleverd waarmee we ons voordeel hebben kunnen doen bij het verder richting geven aan dit project.

Effectmeting

Werk-pakket/ Deliver-able	Gestart ja/nee	Volgens CD of afwijking	Stavaza/evaluatie	Indien afwijking waardoor	Risico
WP1	ja	volgens CD	Afgerond		geen
WP2	Ja	volgens CD	Afgerond		geen
WP3a	ja	volgens CD	Afgerond		geen
WP3b	ja	Volgens CD	Afgerond		geen
WP4a	ja	volgens CD	Afgerond		geen
WP4b	ja	afwijking	Afgerond	Extra pilot-experimenten waren nodig in verband met te kleine deelname in de eerste ronde.	geen
WP4c	ja	afwijking	Afgerond	Extra brainstormsessie nodig op later tijdstip vanwege vakanties (vertraging is niet kritiek)	geen
WP4d	ja	afwijking	Afgerond	In maart 2014 laatste pilot-experiment in Groningen. In de periode maart-juni worden de analyse en rapportage van de resultaten van de pilot-experimenten afgerond.	geen
WP5	ja	afwijking	Afgerond	wordt in periode maart-juni 2014 afgerond in verband met uitloop pilot-experimenten.	geen
WP6a	ja	afwijking	Afgerond	idem	geen
WP6b	ja	volgens CD	Afgerond		geen
WP6c	ja	volgens CD	Afgerond		geen
WP6d	ja	volgens CD	Afgerond		geen

Vooruitblik voorbij de projecthorizon

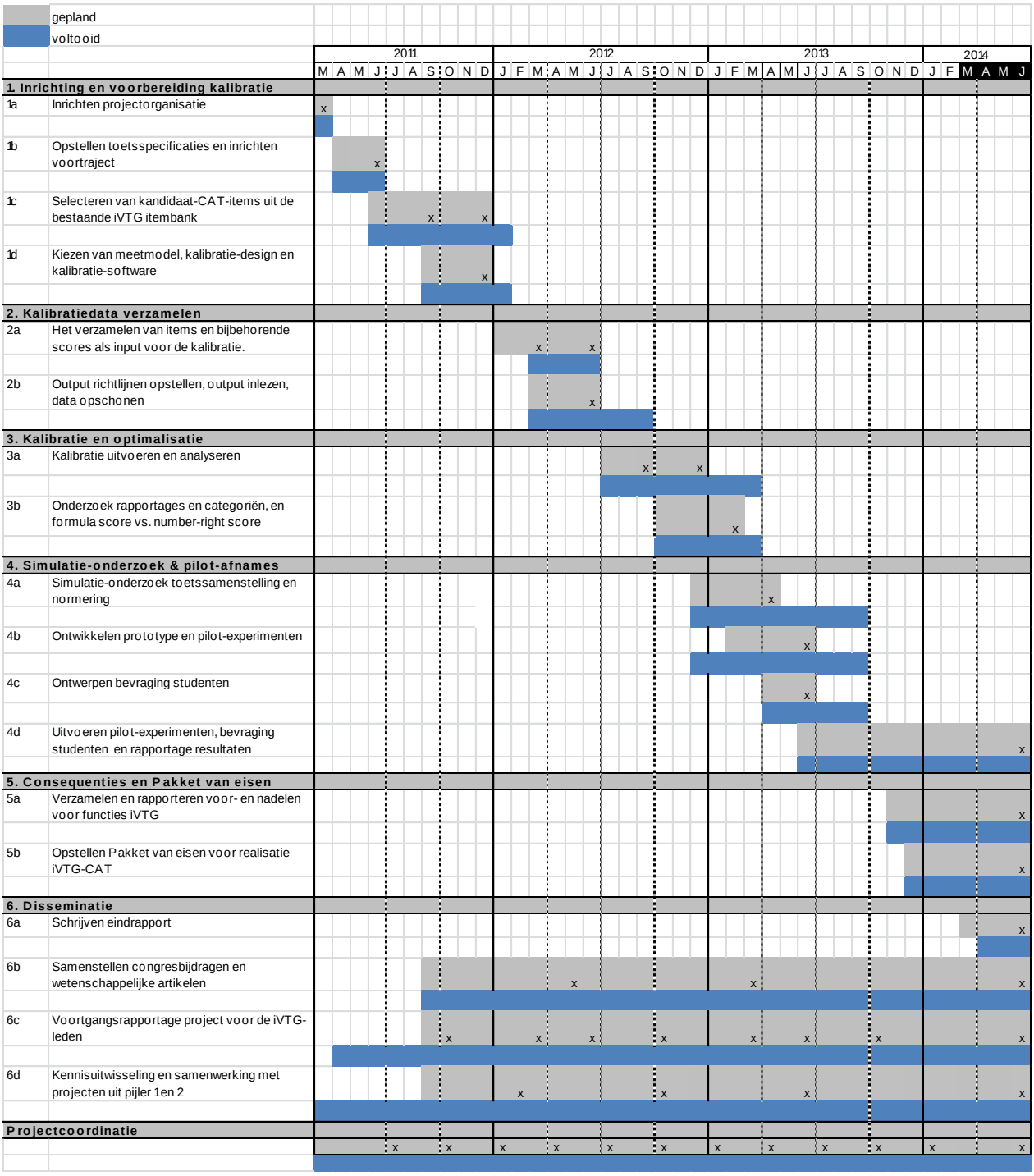
Effectmeting/Exploitatieplan/implementatieplan
AdaPT is een onderzoeksproject naar de haalbaarheid van CAT-iVTG. De aard van het project omvat hierdoor al effectmeting en evaluatie.

Deze aspecten worden besproken in sectie WP5 ‘Consequenties en pakket van eisen’.

Bijlage 1 Financiële rapportage

Volgens afspraak volgt de financiële eindrapportage op 1 oktober 2014.

Bijlage 2 Standlijnenoverzicht



Bijlage 3 Opgeleverde resultaten

Inhoudsopgave

Bijlage1.01_2011_4 Cito-visie op werkplan 1b 1c 1d.docx

Bijlage1.02_WP1B Toetsspecif_Inrichten voortraject_RappCategorien.docx

Bijlage1.03_Rapportage_haalbaarheidsstudie_CAT_geneeskunde_def.doc

Bijlage1.04_Concept handleiding classificatie kennis- en toepassingsvragen.docx

Bijlage1.05_RapportCorrGroeiEnRelev180211.doc

Bijlage1.06_WP1d –kalibratie-design_proefafnames_software_meetmodel.docx

Bijlage1.07_Adapt brainstorm proefafnames.docx

Bijlage1.08_20120307 notitie gebruik historische data 14 febr ADAPT.docx

Bijlage2.01_Script-groei-sprong-rest.docx

Bijlage2.03_Feedback categorieën 14 06 2011.doc

Bijlage3.04_Rapport_JLeppink_kalibratie_Samenvatting van itemkalibratie in OPLM_AMU.docx

Bijlage3.05_Report_Items-Growth-CAT-Itembank.docx

Bijlage4.01_TheoEggen_Prototype adaptieve toets geneeskunde.docx

Bijlage4.02_ProeftoetsQuestify_Memo.docx

Bijlage4.03_ProeftoetsQuestify_10itemsMRT2006.txt

Bijlage4.04_ProeftoetsQuestify_Items-Growth-CAT_vosysids_antw_catdisc_etc.xlsx

Bijlage4.06_00286_ethical_review_c_approve_AMMMuijtjens.pdf

Bijlage4.07_Brief aan Onderwijsdirecteur ivm proefafname AdaPT.docx

Bijlage4.08_Brief aan Bestuur Studiever Pulse ivm proefafname AdaPT.docx

Bijlage4.09_TekstWerving_UM_pilotAdaPT_Nov_2013_MarionStijnen.docx

Bijlage4.10_TekstWerving_UM_pilotAdaPT_Jan_2014_FHMLnieuws_MarionStijnen.docx

Bijlage4.11_TekstWerving_UM_pilotAdaPT_Jan_2014_EleUM_announcement_RogerRennenberg.docx

Bijlage4.12_Informed Consent.pdf

Bijlage4.13_Informatiebrief_CAT.pdf

Bijlage4.14_Instructie ProefafnameExpertUM adaptieve voortgangstoets dd 5 sept 2013.docx

Bijlage4.15_Instructie ProefafnameStudentUM adaptieve voortgangstoets.docx

Bijlage4.16_AdaPT vragenlijst 2013_10_31.docx

Bijlage4.17_Rapport_analyse toetsduur.docx

Bijlage4.18_Rapport_Leppink-ea-AdaPT-CAT-report.docx

Bijlage4.19_Rapport_Analyse_Scores AdaPT vs Vosys_2014_06_30.docx

Bijlage5.01_Notitie_Omvang_Itembank_Operationeel.docx

Bijlage6.01_Programma Cursus Multidimensionele Item Response Theorie.docx

Bijlage6.02_Report of visit to the NBME and MCC.docx

Bijlage6.03_Pres OWD2011_11_09_AMuijtjens_ Adaptieve VT.pdf

Bijlage6.04_Abstr AMEE2012_JVanBerlo_ NL Interuniv ProgrTest in Medicine – furth devel.docx

Bijlage6.05_Abstr AMEE2012_AMuijtjens_Relevance and Growth.docx

Bijlage6.06_Abstr NVMO2012_JVanBerlo_AdaPT_Ontw adaptieve voortg toets geneesk.docx

Bijlage6.08_Abstr NVMO2012_AMuijtjens_Item-groeipatronen en item-relev bij adapt voortg toets.docx

Bijlage6.10_Pres OWD_2012_11_13_AMuijtjens_InterunivVTgeneesk_ErvaringEnRecOntw.pdf

Bijlage6.11_Post_JVanBerlo_SHEacad2013_adapt_poster.ppt

Bijlage6.12_Pres WAS2013_AMuijtjens2013_09-26_WAS_Progress_Testing.pdf

Bijlage6.13_Abstr AMuijtjens_AMEE2014_CAT_Progress_Test_Student_Experiences_AMU260214.docx

Bijlage6.15_Hfdst_MuijtjensA_2014_BookChpt_ToetsenMetVoortgangstoetsing_InVanBerkelToetsenInHO_p169.pdf

Bijlage6.16_Art_MuijtjensA_2013_Voortgangstoetsing_Examens_nr1_p7.pdf

Bijlage6.17_Art_WrigleyW_2012_ASystFramewForTheProgrest_AMEEg71_MedTeach_p683.pdf

Bijlage6.18_Abstr_JeanVanBerlo_A Computerized Adaptive Progres Test_Poster_CAA2014_AMU.docx