

Vergelijkend onderzoek naar Statistische Analyse-Instrumenten

Een vergelijking op grond van geboden functionaliteit en gebruiksvriendelijkheid

Alet Kapma

In dit onderzoek worden vier statistische analyse programma's vergeleken op functionaliteit en gebruiksvriendelijkheid: QMP, Sonate, statistische module van TestVision en WinAnalyse. Deze vergelijking is gemaakt op basis van een inventarisatie- en een gebruikersonderzoek. Het inventarisatieonderzoek bestaat uit twee delen: functionaliteit en gebruiksvriendelijkheid. Om functionaliteit te bepalen is onder andere gekeken naar psychometrische aspecten. Voor het deel gebruiksvriendelijkheid is gekeken welke aspecten van een programma bijdragen aan gebruiksvriendelijkheid. Hier zijn de volgende aspecten uit naar voren gekomen: aantrekkelijkheid, gebruikerscontrole, navigatie, ondersteuning en taal. Deze aspecten zijn ook verwerkt in de vragenlijst, het instrument voor het gebruikersonderzoek. Voor gebruiksvriendelijkheid is er een verschil gevonden tussen Sonate en QMP, waarbij de eerste beter scoorde. Uit het functionaliteitsonderzoek is naar voren gekomen dat TestVision de hoogste functionaliteit biedt. QMP, Sonate en TestVision bieden alle drie voldoende indices om de kwaliteit van toetsen te verbeteren.

Inleiding¹

De laatste jaren is de belangstelling voor de kwaliteit van het onderwijs en daarmee ook voor toetsen toegenomen. Deze belangstelling loopt parallel aan de toename van het gebruik van de computer bij toetsafname. Met de toegenomen mogelijkheden voor het toetsen per computer komt ook de toets- en itemanalyse (TIA) onder handbereik van elke docent. In veel toetssystemen is een statistische analyse module opgenomen. Na toetsafname is de toets- en itemanalyse eenvoudig uit te voeren en na interpretatie van de gegevens wordt duidelijk of de gegeven toets aan de gestelde eis van betrouwbaarheid voldoet en welke items voor verbetering vatbaar zijn.

In het onderzoek zijn de volgende statistische analyse-instrumenten (SAI) opgenomen:

- Question Mark Reporter van QMP (3.4 versie)
- Sonate (1.2 versie)
- Statistische module behorend bij TestVision (4.0.6 versie)
- WinAnalyse als onderdeel van WinToets (3.0 PRO-versie)

De uitkomst van dit onderzoek kan een hulp zijn voor toekomstige gebruikers die een keuze willen maken uit toetsservicesystemen en daarbij ook de mogelijkheden van de bij het toetsservicesysteem behorende SAI betrekken. Ook kunnen ze beslissen om een separate SAI aan te schaffen zoals Sonate.

De vier in dit onderzoek betrokken SAI worden zowel in het bedrijfsleven als in diverse vormen van onderwijs gebruikt. Er is een vergelijking gemaakt tussen deze vier systemen op grond van geboden functionaliteit en gebruiksvriendelijkheid. Voor het maken van deze vergelijking is een inventarisatie van de functionaliteiten van de SAI gemaakt. Om verschillen in gebruiksvriendelijkheid te analyseren is een vragenlijst opgesteld en afgenomen bij de gebruikers van SAI.

De probleemstelling die het startpunt van dit onderzoek vormt luidt:

In welke mate verschillen statistische analyse instrumenten?

Het onderzoek richt zich op de volgende drie aspecten om de verschillen aan te geven. Bij elk aspect is een onderzoeksvraag geformuleerd.

1. Functionaliteiten
Welke functionaliteiten bieden de vier statistische analyse-instrumenten voor het analyseren van toetsen?
2. Kwaliteitsverbetering
Zijn de gegeven indices van de verschillende SAI geschikt voor het verbeteren van de kwaliteit van zowel items als toets?
3. Gebruiksvriendelijkheid
Is er verschil in gebruiksvriendelijkheid van de verschillende analyse-instrumenten?

Statistische analyse-instrumenten

Een SAI is een “computer-based” programma dat een psychometrische analyse uitvoert. Deze psychometrische analyse wordt vaak aangeduid met toets- en itemanalyse (TIA), omdat deze vaak in combinatie uitgevoerd worden. SAI kunnen separaat of als onderdeel van een toetsservicesysteem (TSS) of elektronische leeromgeving (ELO) voorkomen. In een toetsservicesysteem zijn alle componenten samengebracht die nodig zijn voor het afnemen van toetsen. Deze onderdelen zijn volgens Boogaard, Latour & Gremberghe (2003):

- Itemconstructie
- Itembankbeheer
- Toetsontwerp
- Toetssamenstelling
- Toetsafname
- Toetsverwerking en analyse

¹ Met dank aan dr. ir. H.J. Vos van de faculteit Educational Science and Technology aan de Universiteit Twente voor de begeleiding van dit onderzoek. Daarnaast wil ik graag de respondenten en distributeurs van QMP, Sonate, TestVision en WinAnalyse bedanken voor hun medewerking en ook het SVM/NIVO voor het beschikbaar stellen van het enquêteprogramma.

- Toetsevaluatie

Een SAI maakt toetsevaluatie, toetsverwerking en toetsanalyse mogelijk voor iedere docent. De analyse kan uitgevoerd worden op basis van de klassieke testtheorie of de itemresponstheorie. Alle in dit onderzoek opgenomen SAI maken gebruik van de klassieke testtheorie. Tijdens de analyse worden de psychometrische gegevens berekend. De interpretatie van de psychometrische gegevens geeft een indicatie voor de kwaliteit van de items en de toets als geheel. Hierbij wordt er van uit gegaan dat de toets valide is wat betreft inhoud en bruikbaarheid. Als een item niet aan de kwaliteitseisen voldoet kan het item verbeterd worden en daarmee verbetert ook de kwaliteit van een volgende toetsafname. Is de toets eenmaal afgenomen dan kan besloten worden om het item te verwijderen en daarmee de kwaliteit van de afgenomen toets te verbeteren. Een dergelijke beslissing moet altijd in samenhang met de inhoud genomen worden en nooit alleen op grond van de psychometrische gegevens. De inhoudsvaliditeit moet na verwijderen van de items gewaarborgd blijven.

Gebruiksvriendelijkheid

In dit onderzoek wordt onder gebruiksvriendelijkheid verstaan het gemak van het gebruik en de tevredenheid over het programma.

Methode

In dit onderzoek is voor een tweeledige aanpak gekozen. Ten eerste is er een inventarisatieonderzoek uitgevoerd. Hiervoor is een lijst met criteria opgesteld om de SAI te vergelijken. Op de tweede plaats is er een gebruikersonderzoek uitgevoerd. Hiervoor is een vragenlijst afgenomen onder de gebruikers van de vier SAI.

De gegevens van het inventarisatieonderzoek zijn toegestuurd aan de distributeurs van de SAI om eventuele onjuistheden aan te kunnen passen. Hier zijn geen reacties op gekomen.

Instrumenten

Inventarisatie onderzoek

Om de functionaliteiten van de vier SAI te kunnen vergelijken is een criteriumlijst (kruisjestabel) opgesteld. De lijst² is verdeeld in twee delen. Het eerste deel heeft betrekking op de functionaliteit van het programma. Het tweede deel heeft betrekking op de gebruiksvriendelijkheid van het programma.

In een SAI resulteert een toets- en itemanalyse in het algemeen in de gegevens van kandidaat, toets en item. Daarom zijn kandidaat, toets en item als categorie opgenomen in de criteriumlijst. In de lijst zijn hieraan toegevoegd de categorieën instellingen, inlezen diverse vraagtypen en rapportage. De categorie instellingen bevat mogelijkheden die in sommige toetssystemen al aangegeven moeten worden bij het maken van de vragen en het samenstellen van de toets. Strikt genomen vallen ze dan niet bij alle onderzochte SAI binnen de mogelijkheden van het SAI. Voor het tweede deel, waarin de gebruiksvriendelijkheid van het programma beoordeeld wordt, is uitgegaan van de kenmerken die bijdragen aan gebruiksvriendelijkheid van programma's opgesteld door Gaffney (2001). Deze kenmerken zijn:

- Navigatie
- Functionaliteit
- Gebruikerscontrole
- Taal
- Help en ondersteuning
- Feedback
- Consistentie

² De gebruikte lijst is op te vragen bij de onderzoeker

- Fouten
- Aantrekkelijkheid

Van deze kenmerken zijn op grond van inhoud de kenmerken feedback en fouten samengevoegd met help en ondersteuning. Dit om tot een overzichtelijk aantal categorieën te komen met voldoende items per categorie. Taal bevat uiteindelijk maar twee items, maar dit kenmerk kan op grond van inhoud niet samengevoegd worden met één van de andere kenmerken.

De inhoud van de vragen is gebaseerd op deze kenmerken en op de tien vuistregels van Nielsen (1993) die aangeven wat belangrijke aspecten zijn van gebruiksvriendelijkheid van een programma. Deze kunnen een groot deel van de problemen bij het gebruik van computer gestuurde programma's verklaren. De tien vuistregels van Nielsen (1993) zijn:

1. Simpele en natuurlijke dialoog
2. Spreek de taal van de gebruiker
3. Minimaliseer de "memoryload" van de gebruiker
4. Consistentie
5. Feedback
6. Duidelijke "exit"
7. Shortcuts
8. Foutmeldingen
9. Voorkomen van fouten
10. Help en documentatie

De inhoud van deze vuistregels komt op functionaliteiten na overeen met de inhoud van de kenmerken van Gaffney (2001). De verschillende categorieën die zijn opgenomen in de criteriumlijst staan in Tabel 1. Tevens is het aantal items per categorie vermeld.

Tabel 1

Categorieën in de lijst met criteria

Deel A: Functionaliteiten	Aantal items	Deel B: Gebruiksvriendelijkheid	Aantal items
Inlezen diverse vraagtypen	8	Navigatie	9
Instellingen	13	Gebruikers controle	7
Gegevens kandidaat	4	Taal	2
Gegevens toetsanalyse	28	Ondersteuning	11
Gegevens itemanalyse	21	Aantrekkelijkheid	6
Rapportage	8		

Vragenlijst

Er is een vragenlijst opgesteld die bestaat uit twee delen met gesloten vragen: een algemeen deel en een deel met vragen over de waardering van het SAI. De vragen in het algemene deel zijn nodig voor een goede interpretatie van de gegevens. Het aantal alternatieven is afhankelijk van de vraaginhoud en varieert van twee tot vijf. In het tweede deel zijn vijf categorieën onderscheiden die betrekking hebben op de gebruiksvriendelijkheid. Deze categorieën zijn gebaseerd op de kenmerken van Gaffney (2001) en de vuistregels van Nielsen (1993).

De vijf categorieën bevatten elk een aantal stellingen die op een 5-punts Likert-schaal beantwoord worden. Gekozen is voor een 5-puntsschaal omdat dit de gebruiker voldoende keus geeft om zijn mening weer te geven. De antwoord categorieën op deze schaal lopen van 1 tot en met 5, waarbij 1 "helemaal mee oneens" en 5 "helemaal mee eens" weergeeft. Om eventuele onduidelijkheden en onvolledigheden te achterhalen in de vragenlijst is een "pretest" uitgevoerd. Daarvoor zijn een aantal interviews afgenomen. De concept vragenlijst is voorgelegd aan een vijftal gebruikers en een tweetal deskundigen op het gebied van het afnemen van vragenlijsten. Tijdens deze interviews is de vragenlijst doorlopen en zijn vervolgens inhoudelijke en uiterlijke aspecten besproken. Na afloop van deze interviews zijn een aantal wijzigingen in het concept aangebracht.

De volgende overwegingen hebben geleid tot de keuze voor het online afnemen van de vragenlijst:

- Gebruikersgroepen geven e-mail adressen, geen huisadressen of telefoonnummers.
- Interviewen kost veel tijd (Swanborn, 1994). Omdat de vragenlijst hooguit 15 minuten van de onderzoeker vraagt wordt hiermee een grotere respons verwacht dan bij interviewen.
- E-mail werkt snel.
- Het online afnemen van een enquête is goedkoper.

Respondenten

Omdat het een onderzoek naar de gebruiksvriendelijkheid betreft van vier SAI, is gekozen voor een aselechte gestratificeerde steekproef (Swanborn, 1994). Omdat de populatie van gebruikers van de betreffende SAI onbekend is, is gekozen voor “snowball sampling” (Krathwohl, 1998). Eerst zijn de distributeurs van de betreffende SAI benaderd met het verzoek om mailadressen door te geven. In eerste instantie hebben alleen de distributeurs van TestVision en Sonate gereageerd. Dit leverde niet het gewenste aantal respondenten op. Daarom zijn ook de contactpersonen van de opleidingen/bedrijven die QMP gebruiken benaderd, deze contactpersonen zijn te vinden op de site van QMP³. De reacties waren erg positief, maar leverde weinig adressen op omdat de meeste gebruikers niet de SAI gebruiken. Aangezien er ook van de distributeur van WinToets geen reactie is gekomen zijn de adressen van de mensen die zich voor de gebruikersgroep van Wintoets hebben aangemeld gebruikt. Deze zijn te vinden op de site van Wintoets⁴. Met elkaar heeft dit 248 respondenten opgeleverd. Van deze 248 hebben er 83 gereageerd (zie Tabel 2), wat een respons opleverde van ruim van 30%. Veel mensen die de vragenlijst niet hebben ingevuld gaven door middel van e-mail aan dat ze nog geen gebruik maken van een SAI, omdat ze in een beginfase van het gebruik van het toetsysteem zaten.

Tabel 2

Aantal respondenten

SAI	Aantal toegestuurd	Aantal reacties
QMP	62	21
Sonate	9	10
TestVision	18	9
WinAnalyse	159	43
Totaal	248	83

Procedure

Inventarisatieonderzoek

Na analyse van diverse toetsen met de verschillende SAI is deze lijst ingevuld door de onderzoeker.

Gebruikersonderzoek

De ontwikkelde vragenlijst is per e-mail naar gebruikers van SAI gestuurd. Om een maximale respons te krijgen gaat de vragenlijst⁵ vergezeld van een e-mail bericht met een introductie en een retourinstructie (Heuvelmans, 1998). In de introductie worden kort en helder enkele zaken betreffende het onderzoek en de vragenlijst uiteengezet. Na het doorlopen van de vragenlijst wordt de retourinstructie vermeld met een dankwoord en een richtlijn voor het terugsturen van de vragenlijst. Twee weken na het toesturen van de vragenlijst is een herinneringsmail gestuurd om de vragenlijst alsnog in te vullen. Voor het digitaal afnemen van de vragenlijst is het enquêteprogramma “Tevreden” gebruikt, ter **beschikking gesteld door** SVM/NIVO te Nieuwegein.

³ <http://www.perceptiongebruikers.net>

⁴ <http://www.drp.nl/WinToets/gebruikers.htm>

⁵ De gebruikte vragenlijst is op te vragen bij de onderzoeker

Resultaten

Inventarisatie-onderzoek

In het inventarisatie-onderzoek zijn elf categorieën opgenomen. Elke categorie bestaat uit een aantal items. Al deze items kunnen door de onderzoeker beantwoord worden met ja of nee. De score per categorie is berekend door aan elk positief antwoord één punt en aan elk negatief antwoord nul punten toe te kennen. Om de resultaten⁶ te kunnen vergelijken is per SAI de gemiddelde score per categorie berekend. Hieronder staan de belangrijkste verschillen tussen de SAI die uit dit inventarisatie-onderzoek naar voren zijn gekomen.

Inlezen diverse vraagtypen

TestVision en QMP kunnen alle vraagsoorten aan. WinAnalyse ook, met uitzondering van “drag and drop” vragen. Uitzondering is Sonate⁷, die ontwikkeld is voor meerkeuze vragen en geen andere vragen kan verwerken.

Instellingen

Hier komen onderwerpen als kandidaten verwijderen, meerdere antwoorden goed rekenen, instellen van scoringsregel, nauwkeurigheid cijfer wegingsfactor e.d. aan bod.

Hierop scoorden alle programma's matig.

Gegevens kandidaat

QMP geeft geen cijfer aan voor de kandidaat, de andere SAI wel. Een tweede verschil is dat QMP en TestVision aangeven of een kandidaat is geslaagd/gezakt, terwijl Sonate en WinAnalyse dit niet aangeven.

Toetsanalyse

Sonate en TestVision berekenen de meeste indices, gevolgd door QMP. WinAnalyse is zeer beperkt in de mogelijkheden.

Itemanalyse

Ook hier berekent TestVision de meeste indices, gevolgd door Sonate. Op de derde plaats komt QMP, WinAnalyse berekent de minste indices.

Rapportage

Hier zijn geen grote verschillen te zien tussen de verschillende SAI. TestVision scoort gemiddeld net iets hoger dan de overige SAI instrumenten. Sonate biedt de mogelijkheid om de gewenste indices aan te vinken.

Navigatie

De navigatie scoort voor Sonate, TestVision en WinAnalyse ongeveer even hoog. QMP valt iets lager uit vanwege het feit dat de gebruiker niet altijd weet waar hij is.

Gebruikerscontrole

Hier is verschil waar te nemen tussen de SAI. De gebruikerscontrole scoort maximaal bij Sonate. Terwijl WinAnalyse duidelijk lager scoort. QMP en TestVision scoren hier tussenin.

Taal

Het taalgebruik is bij alle onderzochte SAI duidelijk.

Ondersteuning

Hier komen onderwerpen als helpdesk, helpfile etc. aan de orde. QMP, Sonate en TestVision scoorden vergelijkbaar. Door het ontbreken van context gevoelige help valt de score van WinAnalyse iets lager uit.

Aantrekkelijkheid

Hiermee wordt het overzicht en de hoeveelheid informatie per scherm bedoeld. Ook het kleurgebruik binnen het programma kan de aantrekkelijkheid beïnvloeden. Hier zijn duidelijk verschillen waar te nemen tussen de SAI.

⁶ De resultaten zijn op te vragen bij de onderzoeker

⁷ Medio oktober 2004 wordt Sonate geïntegreerd in toetsmanagementsysteem Etude 3.0 en kan in deze vorm alle gangbare vraagsoorten aan.

Niet elke categorie bevat evenveel items. Om de resultaten per categorie te kunnen vergelijken is daarom een gestandaardiseerde score berekend door de behaalde score per categorie te delen door de maximale score. De percentages van deze gestandaardiseerde score zijn weergegeven in Tabel 3.

Tabel 3

Percentage gestandaardiseerde score per categorie en per deel

	QMP	Sonate	TestVision	WinAnalyse
<i>Deel A: Functionaliteiten</i>				
Inlezen diverse vraagtypen	88	38	100	88
Instellingen	31	46	54	38
Gegevens kandidaat	75	75	100	75
Toetsanalyse	39	68	71	29
Itemanalyse	33	52	62	19
Rapportage	50	63	75	50
Totaal functionaliteiten	44	57	71	38
<i>Deel B: Gebruiksvriendelijkheid</i>				
Navigatie	67	78	78	78
Gebruikerscontrole	86	100	86	43
Taal	100	100	100	100
Ondersteuning	63	71	63	63
Aantrekkelijkheid	17	67	100	50
Totaal gebruiksvriendelijkheid	63	77	80	63

Gebruikersonderzoek

Er hebben in totaal 84 personen mee gedaan aan het onderzoek, waarvan 7% vrouwen en 93% mannen. Vier personen zijn buiten het onderzoek gelaten omdat ze de vragenlijst onvolledig ingevuld hadden. De leeftijd van de respondenten valt voor 51% in de categorie ouder dan 45 jaar. De instelling waar de respondenten werkzaam zijn liep nogal uiteen per SAI (zie Tabel 4).

Tabel 4

Verdeling in aantallen respondenten over instellingen

	QMP	Sonate	TestVision	WinAnalyse
Voortgezet onderwijs	3	0	0	38
HBO	4	0	4	1
Universiteit	1	10	1	0
Andere instelling	13	0	4	4

Om de gebruiksvriendelijkheid van de SAI te meten is gekeken naar vijf aspecten: navigatie, gebruikerscontrole, taal, ondersteuning en aantrekkelijkheid. Elke categorie omvat een aantal vragen op de vijfpuntsschaal van Likert, waarvan het gemiddelde is bepaald. Voor elk van deze categorieën is een betrouwbaarheidsanalyse uitgevoerd. Dit had tot resultaat dat uit de categorie ondersteuning vier vragen verwijderd zijn. Uit de lage item-rest correlatie ($< .2$) bleek dat deze vragen niet goed in de categorie pasten. Na gekeken te hebben naar de inhoud van de vragen bleek inderdaad dat deze vragen niet eenduidig te interpreteren waren. Door verwijdering van deze vragen gaat de betrouwbaarheid omhoog. Ook voor de categorie aantrekkelijkheid en controle zijn op dezelfde manier twee vragen uit elk verwijderd. De betrouwbaarheid na het verwijderen van de items is in Tabel 5 aangegeven.

Tabel 5
Betrouwbaarheid per categorie

Categorie	Aantal items	Cronbachs alpha
Aantrekkelijkheid	7	0.87
Gebruikerscontrole	6	0.70
Navigatie	6	0.79
Ondersteuning	8	0.86
Taal	3	0.74

Volgens Spector (1992) moet de betrouwbaarheid minstens 0.7 zijn wil de groep van items hetzelfde meten. Uit deze bovenstaande resultaten mag geconcludeerd worden dat het instrument betrouwbaar is. Vervolgens is gekeken of er verschil is tussen de gemiddelden van de verschillende SAI, voor elk van deze categorieën. Dit is gedaan met de één factor variantie-analyse (One-Way ANOVA). Uit deze analyse met een α van 0.05 blijkt dat er tussen de SAI geen significant verschil in gemiddelden is voor de categorie aantrekkelijkheid ($F(3, 79) = 2.49$, $MSE = 0.37$, $p = .07$) en ondersteuning ($F(3, 79) = 1.83$, $MSE = 0.52$, $p = .15$). Wel zijn er significante verschillen gevonden voor gebruikerscontrole ($F(3, 79) = 5.41$, $MSE = 0.42$, $p = .00$), navigatie ($F(3, 79) = 5.41$, $MSE = 0.36$, $p = .00$) en taal ($F(3, 79) = 3.82$, $MSE = 0.40$, $p = .01$).

Om uit te zoeken welke groepen significant van elkaar verschillen binnen de categorieën gebruikerscontrole, navigatie en taal, is de Bonferroni-toets uitgevoerd. Voor gebruikerscontrole is er een significant verschil gevonden in waardering door QMP en Sonate gebruikers en QMP en WinAnalyse gebruikers. De gebruikers van Sonate scoorden hier significant hoger ($M = 4.04$, $SD = 0.60$) dan de gebruikers van WinAnalyse ($M = 3.93$, $SD = 0.65$) en QMP ($M = 3.30$, $SD = 0.61$).

Voor navigatie is er een significant verschil tussen Sonate en QMP gebruikers gevonden. De Sonate gebruikers scoorden ook hier significant hoger ($M = 4.21$, $SD = 0.71$) dan de gebruikers van QMP ($M = 3.56$, $SD = 0.53$) en TestVision ($M = 3.34$, $SD = 0.87$). Terwijl de WinAnalyse gebruikers ($M = 3.95$, $SD = 0.54$) significant hoger scoorden dan de TestVision gebruikers.

Voor taal scoorden de Sonate gebruikers hoger ($M = 4.53$, $SD = 0.42$) dan de gebruikers van QMP ($M = 3.78$, $SD = 0.53$).

Na de verschillen tussen de programma's per categorie te hebben bekeken, is gekeken of er een verschil is in gebruiksvriendelijkheid tussen de programma's. In Tabel 6 staan de aantallen, gemiddelden en standaarddeviaties per SAI aangegeven.

Tabel 6
Overzicht van de gemiddelde scores op de 5 categorieën

	<u>n</u>	<u>M</u>	<u>SD</u>
QMP	21	3.47	0.45
Sonate	10	4.14	0.57
TestVision	9	3.64	0.70
WinAnalyse	43	3.82	0.51

Om te kijken of er significante verschillen tussen de SAI bestaan is een één factor variantie analyse uitgevoerd. Uit deze analyse met een α van 0.05 blijkt dat ertussen de SAI wel een significant verschil in gemiddelden is: $F(3, 79) = 4.23$, $MSE = 0.26$, $p = .008$. Om uit te zoeken welke groepen significant van elkaar verschillen is de Bonferroni-toets uitgevoerd. Hieruit blijkt dat er een significant verschil is bij een α van 0.05 tussen Sonate en QMP ($MD = 0.67$, $MDE = 0.20$, $p = .008$). Er is een 95% betrouwbaarheidsinterval (BI) opgesteld voor de gemiddelde score op Sonate minus de gemiddelde score op QMP. Dit interval loopt van 0.13 tot 1.21. De andere verschillen zijn niet significant.

Conclusie

Functionaliteiten

Terugkomend op de onderzoeksvraag, *Welke functionaliteiten bieden veel voorkomende statistische analyse-instrumenten voor het analyseren van toetsen en waarin onderscheiden deze statistische analyse-instrumenten zich van elkaar?*, kan op grond van het inventarisatie onderzoek het volgende gezegd worden. Er treden grotere verschillen op in functionaliteit dan in gebruiksvriendelijkheid. De grote verschillen in functionaliteit zijn te vinden in de categorieën toetsanalyse, itemanalyse en het inlezen van diverse vraagtypen. Bij de categorie inlezen diverse vraagtypen is opmerkelijk dat Sonate alleen meerkeuze vragen kan verwerken, dit in tegenstelling tot de andere programma's. In de categorie toetsanalyse vermeldt TestVision de meeste en WinAnalyse de minste indices. Opvallend is het ontbreken van toetsbetrouwbaarheid bij WinAnalyse. Binnen de categorie item-analyse komt de geboden functionaliteit van Sonate en TestVision in grote lijn overeen. Hetzelfde geldt voor QMP en WinAnalyse. Deze laatste twee SAI vermelden minder indices dan Sonate en TestVision. Voor de categorie gegevens kandidaat is het opmerkelijk dat de vermelding gezakt/geslaagd ontbreekt bij Sonate en WinAnalyse.

Kwaliteitsverbetering

De item- en toetsindices mogen niet afzonderlijk geïnterpreteerd worden, maar altijd in samenhang met elkaar en de inhoud. Uit literatuuronderzoek (Kapma, 2004) is gebleken dat voor een kwaliteitsoordeel van een item de volgende 3 indices minimaal gegeven moeten zijn: p-, a- en R_{it} - of R_{it} -waarde. Voor een kwaliteitsoordeel over de toets moet minimaal de betrouwbaarheid bekend zijn. Rekening houdend met bovenstaande aspecten en de uitkomst van het inventarisatieonderzoek kan de tweede onderzoeksvraag beantwoord worden. Deze luidt: *Zijn de gegeven indices van de verschillende SAI geschikt voor het verbeteren van de kwaliteit van zowel item en toets?*

Alle onderzochte SAI berekenen een p- en a-waarde. QMP, Sonate en TestVision berekenen ook een R_{it} -waarde of vergelijkbare waarde, waardoor deze drie analyse instrumenten goed uitgerust zijn om de kwaliteit van een item vast te stellen en te verbeteren. Om de kwaliteit van de toets als geheel te bepalen is de betrouwbaarheid nodig. QMP, Sonate en TestVision berekenen de betrouwbaarheid. WinAnalyse geeft geen betrouwbaarheid van de toets aan en is daardoor minder geschikt om een kwaliteitsoordeel over de toets als geheel te geven.

Van de onderzochte SAI kan samenvattend gezegd worden dat QMP, Sonate en TestVision voldoende informatie geven om de kwaliteit van items en toets vast te stellen waarna eventueel verbetering van kwaliteit kan volgen.

Gebruiksvriendelijkheid

Uit het onderzoek is naar voren gekomen dat de gebruikers van Sonate significant meer tevreden zijn over de gebruiksvriendelijkheid van hun programma dan de andere drie groepen gebruikers. De onderzoeksvraag *“Is er verschil tussen de gebruiksvriendelijkheid van de verschillende analyse-instrumenten?”* kan positief beantwoord worden voor Sonate en QMP. Tussen de andere SAI is geen significant verschil gevonden.

Discussie

Zowel in het inventarisatie- als het gebruikersonderzoek komt Sonate er het meest gebruiksvriendelijk uit. Voor TestVision en WinAnalyse is minder overeenkomst tussen de twee onderzoeken. Om hier meer duidelijkheid over te krijgen zou het goed zijn daar in een vervolgonderzoek op terug te komen. Het uiterste moet dan gedaan worden om een hogere respons te behalen. QMP scoorde vrij laag op het gebruikersonderzoek. Een mogelijke verklaring hiervoor is dat de gebruikers van QMP voor het grootste deel werkzaam zijn bij andere instellingen dan het onderwijs. Mogelijk is deze gebruikersgroep kritischer ten aanzien van gebruiksvriendelijkheid dan de gebruikers binnen het onderwijs. Sonate komt als meest gebruiksvriendelijk uit het onderzoek, daar wil ik wel een kleine kanttekening bij maken.

Deleted: ¶
¶

Via de distributeur heb ik negen adressen gekregen, allemaal medewerkers van de TU Delft, waar ook het programma ontwikkeld is. Er zijn tien reacties van Sonate binnen gekomen. Dit is mogelijk, want door een misverstand hebben alle respondenten, ook de mensen die de vragenlijst al hadden ingevuld, een herinneringsmail gekregen waardoor ze nog een keer de vragenlijst konden invullen.

Uit de interviews en de reacties van instellingen/bedrijven die ik heb ontvangen blijkt dat SAI hoofdzakelijk gebruikt worden voor het efficiënt en effectief vaststellen van een cijfer voor de kandidaat. Er wordt weinig gebruik gemaakt van de mogelijkheid om toetsen te analyseren en vervolgens deze informatie te gebruiken voor kwaliteitsverbetering van toetsen. Dat hier zo weinig gebruikt van gemaakt wordt in een tijd waarin de kwaliteit van onderwijs zo in het centrum van de belangstelling staat is teleurstellend. Misschien is het goed als de distributeurs van de SAI hier meer aandacht aan besteden.

Als laatste wil ik opmerken dat de SAI zijn vergeleken op grond van functionaliteit en gebruiksvriendelijkheid. Een ander belangrijk aspect namelijk de prijs van de programma's is niet opgenomen. WinAnalyse⁸ is veruit het voordeligste programma en richt zich vooral op het voortgezet onderwijs. In veel categorieën doet het niet onder voor de andere SAI. Indien WinAnalyse een toetsbetrouwbaarheid en een R_{it} -waarde aan zou geven zou het een heel voordelig alternatief zijn voor QMP, Sonate en TestVision. WinAnalyse zou dan ook geschikt zijn om de kwaliteit van item en toetsen te verbeteren.

⁸ Najaar 2004 komt versie 3.1 uit met meer mogelijkheden.

Referenties

Boogaard, v.d. G., Latour, I. & Gremberghe, v. P. (2003). Vergelijking toetssystemen: een vergelijkend onderzoek naar elektronische toetssystemen en toetscomponenten van elektronische leeromgevingen. Arnhem: Cito.

Gaffney, G. (2001). Introduction to web usability. Retrieved March 21, 2004, from Information & Design pty. ltd. Melbourne Web site: www.infodesign.com.au

Heuvelmans, A.P.J.M. (1998). Constructie en verwerking van vragenlijsten. Arnhem: Cito.

Kapma, A. (2004). Kwaliteitsverbetering van toetsen: de functie van psychometrie in toetsanalyse. Enschede: Universiteit Twente.

Krathwohl, D.R. (1998). Educational & Social Science Research: an integrated approach. New York: Addison-Wesley Educational Publishers.

Nielsen, J. (1993). Usability engineering. London: Academic Press Limited.

Spector, P.E. (1992). Summated rating scale construction: an introduction. London/New Delhi: Sage Publications.

Swanborn, P.G. (1994). Methoden van sociaal-wetenschappelijk onderzoek. Meppel: Boom.