SAMENVATTING

De testen uit Muiswerk Testsuite 7 Nederlands 1F-2F-3F-4F zijn genormeerd met behulp van de ankertesten van het Ministerie van Onderwijs, Cultuur en Wetenschap. Een groot aantal leerlingen heeft beide testen gemaakt. Uit de resultaten is gebleken dat de cesuur van de Muiswerktesten 65% is. Dat is de norm waarop kandidaten het betreffende niveau gehaald hebben. Tevens is aangetoond dat de Drempeltoetsen Taal van Muiswerk voldoende betrouwbaar en valide zijn.

HET ONDERZOEK

De Commissie voor Toetsen en Examens van het Ministerie van Onderwijs, Cultuur en Wetenschap heeft in september 2014 sets referentietesten Lezen en begin 2015 sets referentietesten Taalverzorging ter beschikking gesteld voor uitgevers van testmateriaal. Deze referentietesten, ook wel ankertesten genoemd, zijn bedoeld om bij elk test de norm (cesuur) te bepalen waarop een kandidaat het niveau wel of niet gehaald heeft.

In de periode september 2015 tot juni 2016 zijn de Ankeronderzoeken t.b.v. Muiswerk Testsuite 7 Nederlands 1F-2F-3F-4F uitgevoerd op een aantal scholen.

De Muiswerk testen bevatten voor elk niveau 84 vragen. Uit de set anker vragen zijn per niveau 55 vragen geselecteerd, 20 voor Tekstbegrip, 25 voor Spelling en 10 voor Grammatica. Met deze vragen is een nieuwe module voor Muiswerk gemaakt en deze is vervolgens in een speciale, niet openbare, omgeving ondergebracht.
Er zijn 945 kandidaten getest op het 1F-niveau en 303 kandidaten op het 2F-niveau.

Het eerste doel van dit onderzoek is de norm (cesuur) bepalen waarop bij de Drempeltoetsen Taal van Muiswerk kandidaten het betreffende niveau al dan niet gehaald hebben.

Omdat het hier gaat om het bepalen van een norm en er geen factoren zijn die het testresultaat beïnvloeden, anders dan een gebrek aan kennis en/of vaardigheid van de kandidaat, kan volstaan worden met onderzoeken in hoeverre beide testen als ‘parallel’ beschouwd kunnen worden.

In het boek Testtheorie van prof. Drenth[i] wordt bewezen dat twee testen parallel zijn indien hun gemiddelden gelijk zijn, hun varianties gelijk zijn en de correlatie van beide testen met een onafhankelijke parameter gelijk is. Dat hebben we kunnen bewijzen. Omdat de betrouwbaarheid van de ankertesten hoog is, is de correlatie tussen ankertest en de Muiswerktest voldoende om de betrouwbaarheid van de Muiswerk test te bepalen en hoeven we geen schattingen te doen voor bv een Cronbach’s α.

Twee testen zijn ook parallel als hun gemiddelden afwijken, maar aan de overige criteria wel voldaan wordt. Zij kunnen dan niet zonder meer voor elkaar ingewisseld worden zonder met de specifieke cesuur rekening te houden. Een A- en B-test zou wel eenzelfde cesuur moeten hebben, maar een ijkingsinstrument zoals de hier gebruikte ankertesten hoeven dat niet.

Het bepalen van de cesuur kan op verschillende manieren gedaan worden. Allereerst wordt de cesuur vanuit de ankerset op een lineaire manier berekend. Dat wil zeggen dat de cesuur bepaald wordt uit het gemiddelde van alle vragen. In de praktijk wordt de uitslag van de test bepaald door het gemiddelde van Tekstbegrip en Taalverzorging, waarbij Taalverzorging weer het gemiddelde van

Spelling algemeen, Spelling werkwoorden, Spelling leestekens/interpunctie en Grammatica is. De Muiswerktesten bevatten bovendien nog het onderdeel Luisteren.

Bij het overbrengen van de cesuur zijn alle mogelijke combinaties gemaakt. Doordat het gebruik van de Muiswerktesten gebaseerd is op de syllabus en dus op het bepalen van de gemiddelde score uit de domeinen is ervoor gekozen om de cesuur te bepalen op basis van de twee hoofddomeinen Tekstbegrip en Taalvaardigheid. Daarvoor is de cesuur van de ankerset herrekend naar een cesuur van het gemiddelde van deze domeinen en die is vervolgens vergeleken met de scores van de drie domeinen in de Muiswerktest. Verder zijn ook andere vergelijkingen gemaakt, maar die leverden geen significante verschillen op.

Bij de ankertesten zijn gegevens en hulpmiddelen aangeleverd om de cesuur te bepalen. Voor de 1F-ankertest is de cesuur voor de gekozen vragen vastgesteld op 56,7%. Uit de testresultaten blijkt dat een score van 56% op de ankertesten overeenkomt met een score van 59,2% op de Muiswerktest met een standaarddeviatie van 9,9%. Dat betekent dat de cesuur van de Muiswerk 1F-test op 60% gelegd kan worden. De correlatie tussen de beide 1F testen is 0,50. Daarmee is de Muiswerk Testsuite 7 Taaltest 1F voldoende betrouwbaar om ingezet te worden voor het bepalen van het wel of niet halen van het 1F-taal niveau.

Bij de itemselectie voor de 2F-test blijkt de cesuur bij 61,8% te liggen. Uit de testresultaten blijkt dat een score van 61,8% op de ankertest overeenkomt met een score van 73,1% op de Muiswerktest met een standaard deviatie van 6,4%. Dus bij de Muiswerk-2F-test ligt de feitelijke cesuur op 73%. De correlatie tussen de ankertest 2F en de Muiswerktest 2F is 0,60. Daarmee is de Muiswerk Testsuite 7 Taaltest 2F voldoende betrouwbaar om ingezet te worden voor het bepalen van het wel of niet halen van het 2F-taalniveau. Door de hoge betrouwbaarheid en het feit dat de ankertesten valide zijn, is tevens aangetoond dat de testen van Muiswerk Testsuite 7 Nederlands 1F-2F-3F-4F valide zijn.

Omdat het technisch onhandig is om binnen Muiswerk met verschillende cesuren voor elke test te werken is er voor gekozen om de cesuur steeds op 65% te leggen. Dat betekent dat de 1F-test iets zwaarder beoordeeld wordt en de 2F-test iets lichter.

Oorzaken van verschil in testresultaten

Mogelijke oorzaken voor de verschillen in testresultaten tussen de Ankertest en de Muiswerk Drempeltoetsen taal kan verklaard worden uit de aard van van Tekstbegrip en de gekozen aanpak voor Spelling en de algemene structuur van de Muiswerktesten. Begrijpend lezen is een veel minder eenduidige vaardigheid dan bijvoorbeeld spelling of grammatica. Leerlingen moeten niet alleen details uit een tekst kunnen halen, maar ook (onder andere) gegevens met elkaar in verband brengen, onderwerp en hoofdgedachte bepalen en vaststellen wat de bedoeling van de schrijver is. Het proces vereist integratie van een groot aantal informatiebronnen, van kennis van woordbetekenissen tot kennis van de wereld en vraagt ook nogal wat van concentratievermogen en geheugen. Doordat begrijpend lezen zo'n complex gebeuren is, is het misschien wel onmogelijk om twee tekstbegriptoetsen in hoge mate vergelijkbaar te maken. Het onderdeel Spelling van de Ankertesten bestaat uit productieve dicteevragen en metatalige meerkeuzevragen over de spelling van woorden. De Muiswerktest bevat alleen productieve dicteevragen bij zowel Spelling Algemeen als Werkwoordspelling. Ten slotte is bij de

Muiswerk Drempeltoetsen een bepaalde mate van adaptiviteit ingebouwd. De algemene structuur van de Muiswerktesten bestaat per onderdeel uit vragen van 3 niveaus. Ingestoken wordt op een niveau van gemiddelde moeilijkheid. Met het resultaat daarvan wordt vervolgens gekozen voor een gemakkelijker niveau met afstraffing van punten of voor een moeilijker niveau met het belonen van punten voor het gemakkelijke niveau. Ook in dit mechanisme kunnen verstorende elementen zitten, waarvan de uitwerking niet verder onderzocht is.

Detail uitwerking

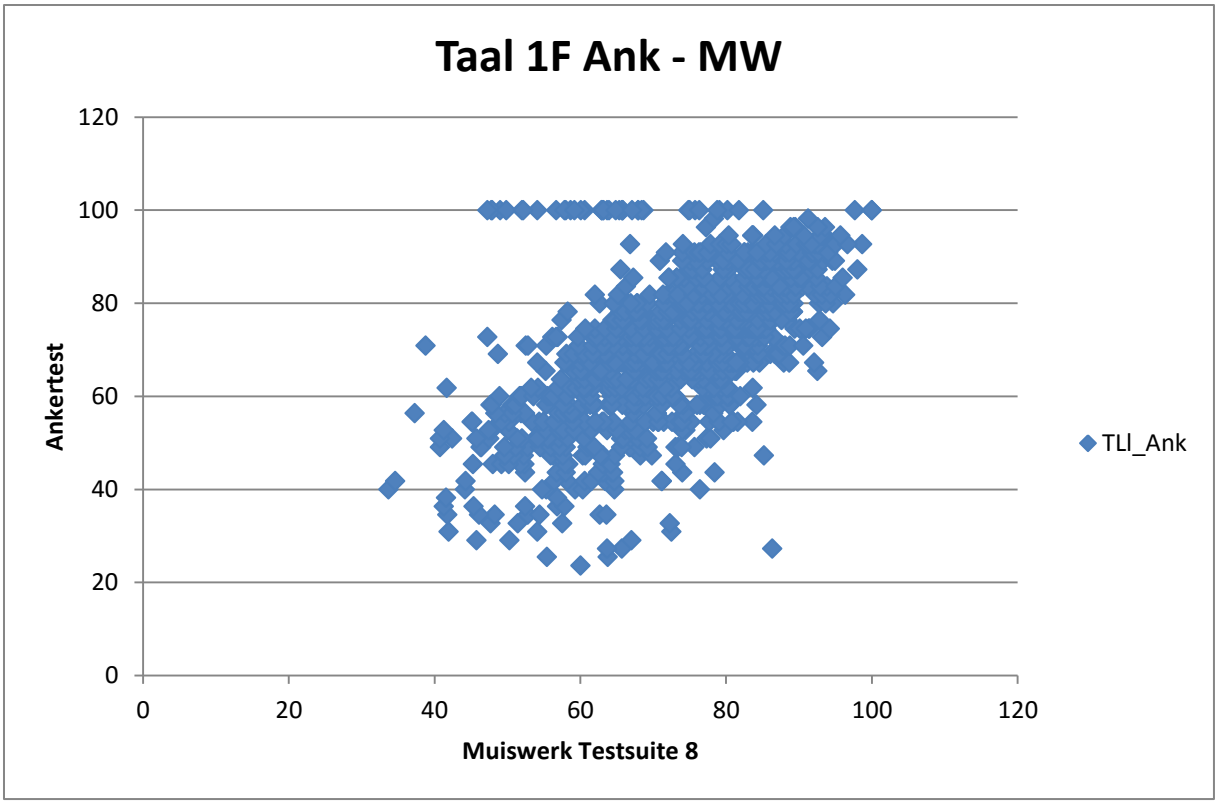
Volgens prof. dr. P.J.D. Drenth en prof. dr. K Sijtsma in Testtheorie (2006)[i] zijn twee testen parallel als aan 3 voorwaarden voldaan is: [1] het gemiddelde is gelijk, [2] de varianties zijn gelijk en [3] de correlaties met een willekeurige onafhankelijke variabele is gelijk, zie bladzijde 209 van bovengenoemd boek.

Hierbij valt op te merken dat Drenth en Sijtsma de derde eis het zwaarst vinden wegen. Wij merken verder op dat het gemiddelde niet gelijk hoeft te zijn, omdat het er juist om ging de cesuur te bepalen. Het zijn dus niet zuiver parallelle testen, maar in principe verschoven parallelle testen.

Hieronder noemen we de Muiswerktest de F-test en de Ankertest de A-test.

Voor de 1F test hebben wij de volgende waarden gevonden.
Bij een score van 56,7% op de A-test is het gemiddelde van de F-test 59,20% met een standaarddeviatie van 9,94%. De variantie van de F-test is 140 en van de A-test 261.
Als onafhankelijke variabele L hebben wij de datum van de eerste testafname genomen.
De correlatie tussen F en L is $-8{,}039 \cdot 10^{-3}$ en de correlatie tussen A en L is $-1{,}108 \cdot 10^{-2}$. In termen van correlatie liggen deze voldoende dicht bij elkaar. Alleen de variantie wijkt enigszins af. Zie de discussie hierboven over de oorzaken van het verschil in testresultaten.

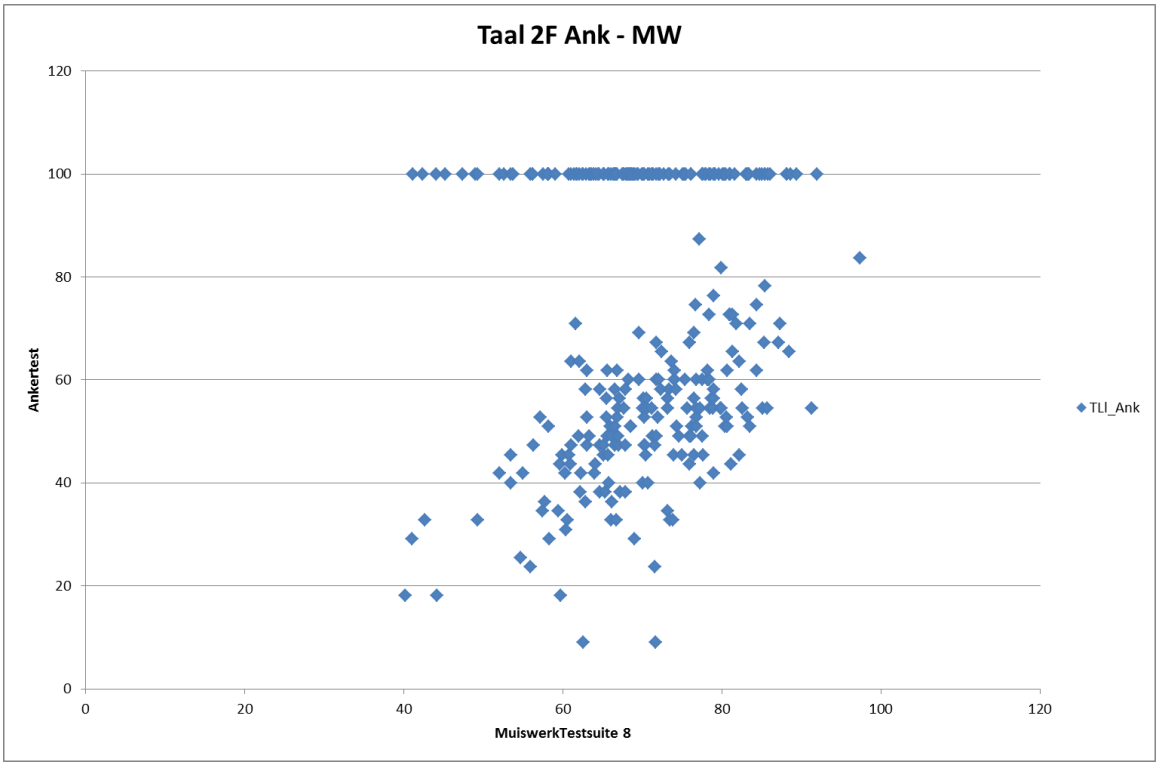
Hieronder het scatterdiagram van de twee testen, waarbij de resultaten van de Ankertest lineair en die van de Muiswerktest via domeingemiddelden berekend is.



Voor de 2F-test hebben wij de volgende waarden gevonden.
Bij een score van 61,8% op de A-test is het gemiddelde van de F-test 73,08% met een standaarddeviatie van 6,37%.
De variantie van de F-test is 396 en van de A-test 847.
Als onafhankelijke variabele L hebben wij weer de datum van de eerste testafname genomen.
De correlatie tussen F en L is $-1{,}04 \cdot 10^{-3}$ en de correlatie tussen A en L is $+6{,}26 \cdot 10^{-3}$. De waarden liggen voldoende dicht bij elkaar.

Wel wijkt ook hier de variantie enigszins af, zie de eerdere discussie.

Hieronder ziet u het scatterdiagram van de twee testen, waarbij de resultaten van de Ankertest lineair en van de Muiswerktest via domeingemiddelden berekend is.



Formules

Gebruikte formules:

Variantie: $S^2(F) = \frac{1}{n}\sum_{i=1}^{n}(F_i - \overline{F})^2$

Covariantie: $V(F,L) = \frac{1}{n}\sum_{i=1}^{n}(F_i - \overline{F})(L_i - \overline{L})$

Correlatie: $r(F,L) = \frac{\frac{1}{n}\sum_{i=1}^{n}(F_i - \overline{F})(L_i - \overline{L})}{S(F)S(L)}$

Ir. Freek W.Weeda Delta-Plus Training & Consultancy www.delta-plus.nl Culemborg, juli 2016

[i] Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie*. Houten: Bohn Stafleu van Loghum.